International Journal of Medical Informatics 63 (2001) 41–50

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

Feature mining and predictive model construction from severe trauma patient's data

Janez Demšar [a,*], Blaž Zupan [a,b,c], Noriaki Aoki [c,d], Matthew J. Wall [e], Thomas H. Granchi [e], J. Robert Beck [c]

[a] *Faculty of Computer and Information Sciences, University of Ljubljana, Tr·aška 25, SI-1000 Ljubljana, Slovenia*
[b] *J. Stefan Institute, Ljubljana, Slovenia*
[c] *Office of Information Technology, Baylor College of Medicine, Houston, TX, USA*
[d] *Department of General Medicine and Clinical Epidemiology, Kyoto University, Kyoto, Japan*
[e] *Ben Taub General Hospital, Houston, TX, USA*

Abstract

In management of severe trauma patients, trauma surgeons need to decide which patients are eligible for damage control. Such decision may be supported by utilizing models that predict the patient's outcome. The study described in this paper investigates the possibility to construct patient outcome prediction models from retrospective patient's data at the end of initial damage control surgery by using feature mining and machine learning techniques. As the data used comprises rather excessive number of features, special attention was paid to the problem of selecting only the most relevant features. We show that a small subset of features may carry enough information to construct reasonably accurate prognostic models. Furthermore, the techniques used in our study identified two factors, namely the pH value when admitted to ICU and the worst partial active thromboplastin time, to be of highest importance for prediction. This finding is pathophysiologically reasonable and represents two of three major problems with severe trauma patients, metabolic acidosis, hypothermia, and coagulopathy. © 2001 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Severe traumatic injury; Damage control; Data mining; Feature mining; Machine learning; Medical prognostic models

1. Introduction

Recent advances in machine learning, data mining and intelligent data analysis has resulted in increased utility of their methods to derive medical prognostic models from retrospective data [8,10,14]. On one side, this can contribute to increased availability and volume of medical data gathered through systematic use of laboratory, clinical and hospital information systems. On the other

[*] Corresponding author. Tel.: +386-1-4768-386.
E-mail address: janez.demsar@fri.uni-lj.si (J. Demšar).

PII: S1386-5056(01)00170-8

side, the aforementioned modeling techniques have matured and may, beside the ability to construct highly predictive models, support feature mining through informed selection and transformation of most relevant features from the data [5,7], construction of interpretable prognostic models [8], handling of noise and missing values, and discovery and incorporation of non-linear patterns and feature combinations.

The paper investigates the utility of feature mining and machine learning techniques to construct an outcome prediction model for severe trauma patients after the first surgery. Trauma surgeons face complex management and decision making problems when treating patients with severe traumatic injury. In the initial period of treatment, the patient's continued hemodynamic instability may increase the risk of difficulty of definitive repair of all injuries. Bold attempts to completely correct acute surgical problems, especially trauma, were explored in 1970s and 1980s. The surgical goals of extensive reconstruction and resection at the initial operation were achieved, but the patients went on to die of respiratory failure, multiple organ distress, and coagulopathy.

The damage control approach emerged from a need to meet the challenge of the changing scope and severity of injury. The basic concept of damage control for trauma patients is to avoid extensive procedures on unstable patients, stabilizing fatal problems at initial operation, and applying staged surgery after successful initial resuscitation. Damage control, however, requires a massive investment of personnel, efforts, and resources in a small group of critically injured patients who carry a mortality rate in excess of 50%, even under the best circumstances. From the viewpoint of resource allocation, a reliable prognostic model at the end of initial damage control surgery is desired to optimize the use of limited medical resources.

To develop a corresponding outcome prognostic model, a particular problem addressed in our study was that the database of patient records from Ben-Taub General Hospital in Houston used included rather excessive number of features, so a special attention was paid to the problem of selecting only the most relevant ones. Although the number of patient records in the data set was relatively small, we show that a small subset of features may carry enough information to construct reasonably accurate prediction models.

The paper is organized as follows. Section 2 introduces the dataset that was used to investigate the plausibility of modeling the outcome for severe trauma patients. The feature mining, machine learning and model evaluation methods are introduced in Section 3. The results of data analysis, construction of predictive models and their evaluation are reported in Section 4. Section 5 summarizes the results and concludes the paper.

2. Data

We examined 68 patients retrospectively who required damage control surgery at Trauma and Critical Care Center, Ben-Taub General Hospital, Houston, TX, in the period from 1994 to 1997. A set of 174 features including patient characteristics, features of prehospital care, physical and laboratory findings in emergency room, operating room and intensive care unit (ICU) was used in the analysis. The data set included many missing values; preliminary data set inspection showed that for 78 features data was missing for at least 50% of patients—these features were not included in further analysis. The resulting data set (68 patients, 96 features) had 20.7% of missing values. Out of sixty-eight patients, 45 (66.2%) have died during their stay at hospital.

3. Methods

A number of preprocessing, modeling and performance estimation methods were used in this work. We first describe feature mining techniques that were used to narrow a list of features used. From the resulting data set, prediction models were derived by classification trees and naive Bayes techniques. The performance of the models was assessed through using various criteria and statistical tests.

3.1. Feature mining

Feature mining is a data mining preprocessing stage where, for classification tasks, a subset of most relevant features is identified and potentially reformulated [7]. The identification of most relevant features is most often related to their ranking, subset selection and to their categorization.

In the first step of the preprocessing, we categorized (discretized) the continuous features. This was required for naive Bayes modeling technique, which does not directly handle continuous features. Besides, the mere information on how the features were categorized can be interesting for the domain expert to verify the relevance of the data base (if categorization is as expected) or to point out for new and interesting categories and cut-off points. We have used two approaches for categorization, quartiles and entropy-MDL based discretization. The quartile discretization splits the range of feature values into four intervals, so that the number of patients within each interval is approximately equal. The more sophisticated entropy discretization [2] uses a top-down approach, similar to clustering methods. It starts with an interval covering all the feature values and finds a cut-off point, which maximizes the informativity. Informativity [12] is measured with respect to the outcome; the better the categorized feature can be used to predict the outcome, the higher the informativity. If the gained information is greater than the increase of the minimal description length for the feature values, the interval is cut into two subintervals and the procedure is repeated on both of them. However, it often happens that the process stops at the first step already. In this case, that is, when no useful categorization was found, such feature is regarded as irrelevant. In this way the entropy-based discretization can also be used as a feature selection tool.

As the quartile discretization considers only the values of the feature that is being discretized independently of other features or outcomes, it tends to be more noise-proof on one side but potentially less interesting for the domain expert on the other side. Besides, the number of intervals for the quartile discretization is fixed, so it cannot be used for the feature subset selection.

After categorization, features were ranked using RELIEFF [3,4], which measures usefulness of a feature by observing the relation between its value and patient's outcome. Intuitively, if there is a group of patients with similar feature values, the observed feature is 'valuable' as a predictor if it has different values on pairs of patients with different outcomes (thus distinguishing between them), but the same value on pairs with the same outcome. Features with negative RELIEFF estimate may be considered to be irrelevant. Features with the highest score are presumed to be the most useful for predicting the outcome. In our study, features were ordered according to their RELIEFF scores and presented to the expert who performed the final selection.

3.2. Data modeling

After we have reformulated the trauma patients' descriptions by categorizing and selecting the features, we used two well-known machine learning techniques to induce the predictive models. The first one was our own implementation of *classification trees* derived from a commonly-known ID3 recursive partitioning algorithm [12]. The basic idea of ID3 is to partition the patients into ever smaller groups until creating the groups with all patients corresponding to the same class (e.g. survives, does not survive). To avoid overfitting, we have used a simple pruning criterion that stops the induction when the sample size for a node falls under the prescribed number of examples or when a sufficient proportion of a subgroup has the same output.

The second machine learning method used was a *naive Bayes classifier*. Assuming the independence of predictive variables, the probability that a patient described with values of predictor variables $V = (v_1, \ldots, v_n)$ survives can be estimated by naive Bayes formula [6]

$$P(R|V) = P(R)\prod_{i=1}^{n} \frac{P(R|v_i)}{P(R)}$$

where $P(R)$ is the apriori probability of survival and $P(R|v_i)$ is the conditional probability of survival if ith predictor variable has the value v_i; both are estimated from the training set of patients. The naive Bayes formula used in the paper is correct. Note that this formula can be derived from the more common form

$$P(R|V) = \frac{P(R)}{P(V)}\prod_{i=1}^{n} P(v_i|R)$$

by reusing the Bayes rule

$$P(v_i|R) = \frac{P(R|v_i)P(v_i)}{P(R)}$$

Naive Bayes classifier and classification trees were chosen because they represented two essentially different approaches for induction of predictive models. Naive Bayes models include all of predictive variables used in the data, while classification trees in general only use a subset of most informative features. Naive Bayes models are in essence linear, while classification trees may include more complex relationships. For modeling from medical data, however, it was observed that naive Bayes most often performs best, outscoring classification trees, rules, and even artificial neural networks [1,6].

A baseline for comparison with above two methods was a *majority classifier* that uses a training set to determine the most frequent class and then classifies all cases from the test set to that class.

3.3. Model evaluation methods, metrics, and comparison statistics

After categorizing and selecting the features and inducing outcome prediction models, different statistical measures can be used to estimate the quality of derived models. From those, which we used in this study, the first three (classification accuracy, sensitivity and specificity) consider the class prediction while the other two (average probability assigned to correct class, area under ROC curve) use the model to predict the probabilities of classes.

- **Classification accuracy** (CA) measures the proportion of correctly classified test examples, therefore, estimating the probability of the correct classification.
- **Sensitivity and specificity** (Sens/Spec) measure the model's ability to 'recognize' the

patients of a certain group. If we decide to observe the surviving patients, *sensitivity* is a probability that a patient who has survived is also classified as surviving, and*specificity* is a probability that a not-surviving patient is classified as not-surviving.

- **Average probability assigned to the correct class** (AP) is related to classification accuracy, but it gives additional information on the reliability of the classifier's decisions. If this measure is low, the classifier can still have a good classification accuracy but its decisions are, on the average, marginal.
- **Area under ROC curve** (aROC) is based on a non-parametric statistical sign test and estimates a probability that for a pair of patients of which one has survived and the other has not, the surviving patient is given a greater probability of survival. This probability was estimated from the test data using relative frequencies.

The above metrics and statistics were assessed through stratified *ten-fold cross-validation* [11]. This divides the patient's data set to ten sets of approximately equal size and equal distribution of outcomes. In each experiment, a single set is used for testing the classifier that has been developed from the remaining nine sets. The statistics for each method are then assessed as an average of ten experiments. The same training and testing data sets were used for all classification methods.

The described statistics estimate the quality of a single classifier. Although they can be used to compare classifiers, a better and more statistically correct test is available for this purpose. McNemar's test compares two classifiers by counting the examples, which were classified correctly by the first but not by the second classifier (n_{10}) and vice versa (n_{01}). As the same training and test sets are used for both induction methods, counts can be summed for all ten cross-validation experiments. Under the null hypothesis, the classifiers are equal and so are the counts, $n_{10} = n_{01}$.

The statistics D, computed as

$$D = \frac{(|n_{01} - n_{10}| - 1)}{n_{01} + n_{10}}$$

is distributed approximately by the χ^2 distribution with one degree of freedom.

Another important evaluation of the induced model is done by the domain expert who ultimately decides whether the models make sense and can be of practical prognostic value.

4. Feature mining and model construction

From the set of 96 features, the entropy based discretization found 56 features as irrelevant. RELIEFF assigned negative score to additional four features, thus resulting in a data set with only 36 features. From these, the expert (a board certified emergency physician) selected ten predictive features (features 1–10 in Table 1) considering also their potential clinical significance. The expert additionally verified and confirmed that among features not included in the set of 36 there are none that should be additionally selected for modeling. This confirms the usefulness of feature subset selection in our setting.

The expert also inspected the categorization found by the entropy-based algorithm by using previous reports, pathophysiological interpretation and the additional statistical analysis. For instance, for APPT_WORST he proposed 80 as a simpler boundary than 78.7. For BE_ICU he proposed a higher range of 22.5. These are the only two features that should always be treated as categorical, while other continuous features can also be modeled as continuous, if the modeling technique allows it. The remaining features should be, in his opinion, categorized to three rather than two intervals. We can expect that the method used would indeed devise a finer categorization if more patients were available. Apart from this, the proposed categorization seemed clinically reasonable.

The selected ten features, together with the two-valued outcome (Death, Well) constituted our first data set. Additionally, the expert proposed another feature subset in which MBP_WORST and PH_WORST were replaced by SBP_WORST and PH_ICU (features 11 and 12 in Table 1), respectively.

Modeling algorithms were successful on both data sets. The classification accuracy was especially high for the second one, reaching accuracy of 93% of correct classifications. The conditional probabilities in the naive Bayes classifier and the graphical presentation of the classification tree revealed the models' main strategy; for classification trees, all patients which were given Catecholam were classified to 'Death' and similarly, the conditional probability of survival after being given Catecholam was 0.00. The inspection of the data indeed proved that from 68 patients, all the 16 patients who were given Catecholam died. The expert confirmed that the relation found is sensible but useless. As this drug is usually the last resort used for the most severe patients, it is highly correlated to the patient's outcome but the surgeon cannot use it for making predictions. The expert proposed to remove this feature from the data set for the further experiments.

We, therefore, formed a third data set, with the same features as the second one but with CATECHOLAM removed. The results on this data set are presented in Table 2. Classification trees and naive Bayes classifier are better than the baseline majority classifier, though the statistical significance of the differences is (at best) marginal, probably also due to the low number of patients. Using McNemar's test, the classification tree model with entropy-based discretization was found to be significantly better than majority classifier ($P = 0.04$), while the tree models with quartiles discretization and naive Bayes models with quartiles and with entropy-based

Table 1
Selected features and their description (in alphabetical order)

#	Feature	Categories	Description	Reference
1	APPT_WORST	<78.7, ≥ 78.7	The worst partial active thromboplastin time	25–33 s
2	BE_ICU	<-12.6, -12.6	>Bicarbonate excess at ICU	-2 to 2
3	BLEEDING_T	Yes, No	Physician's impression regarding coagulopathy during operation	No
4	CATECHOLAM	Yes, No	Cathecholamine administration	No
5	EBL	<2.5, ≥ 2.5	Estimated blood loss	
6	MBP_WORST	<36.3, ≥ 36.3	The worst mean blood pressure	>60 mmHg
7	PACO2_OR	<44.0, ≥ 44.0	The worst arterial carbon dioxide tension	35–45 Torr
8	PH_WORST	<7.0, ≥ 7.0	The worst pH	7.35–7.45
9	PT_ICU	<22.3, ≥ 22.3	Prothrombin time at ICU	10.7–13.0 s
10	TYPE_OF_CL	Skin, Bag	The type of closing	
12	PH_ICU	<7.20, 7.20–7.33, >7.33	The worst pH value at ICU	7.35–7.45
11	SBP_WORST	<57.0, ≥ 57.0	The worst systolic blood pressure	>90 mmHg

Table 2
Classification accuracy (CA), average probability assigned to the correct class (AP), sensitivity (Sens) and specificity (Spec) and area under ROC curve (aROC)

Prognostic model	CA	AP	Sens	Spec	aROC
Majority	0.662	0.552	1.000	0.000	0.500
Classification tree (quartiles)	0.824	0.663	0.800	0.870	0.834
Classification tree (entropy)	0.824	0.686	0.822	0.696	0.849
Naive Bayes (quartiles)	0.809	0.777	0.800	0.826	0.891
Naive Bayes (entropy)	0.794	0.777	0.800	0.826	0.882

discretization have significance levels of 0.06, 0.09 and 0.14, respectively.

Fig. 1 shows a classification tree build from the data set with all 68 patients. Notice that because of missing values in the data several patients do not appear in the leaves of the tree. In the data set of 68 patients PH_ICU is defined for 51 patients, and of 20 patients with PH_ICU between 7.20 and 7.33, 18 patients have a defined value for APPT_WORST.

The classification tree was obtained using a simple prepruning (requiring at least two examples in each leaf, and allowing a maximal proportion of 90% of majority class in each internal node). From the expert's perspective, this classification tree is a reasonable model for outcome prediction. It is based on the important representatives from two of the most important groups of factors, which affect the outcome, coagulopathy and acidosis. It is also interesting that the particular importance of this two features to the patient's outcome was theoretically stressed in the work of Rotondo et al. [13]. Actually, the authors claim that the two mentioned features, together with body temperature, are the three that best determine the patient's outcome. Based on our data set, the temperature was left out as not being highly relevant and was excluded already in the categorization phase. Essentially, the reason why other features were estimated as more relevant can be observed from Fig. 2. It compares differences between outcome probabilities for different values of features. The worst body temperature (T_W) and body temperature when the patient was admitted to ICU (T_ICU) are inspected. We used a cut-point at 34 °C for both values. The patients whose body temperature was below 34 °C showed higher mortality compared with those with temperature above 34 °C (51 vs. 23%), however, the difference was not significant, probably due to a small sample size. The differences of outcome probabilities for T_ICU (51 vs. 23%) and for T_W (57 vs. 33%) are smaller than for partial active thromboplastin time (63 vs. 13%), indicating that the latter may have much higher predictive value. Similar was observed when body temperature was compared with other selected features from Table 1 (Fig. 2 shows only the features used in the classification tree from Fig. 1).

The topmost decision in the classification tree from Fig. 1 is based on the blood's pH

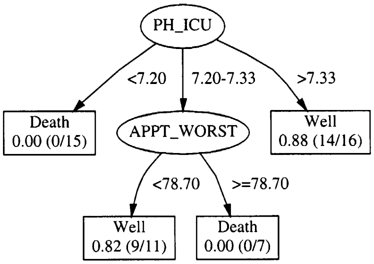

Fig. 1. A classification tree model, derived with entropy discretization with simple prepruning.

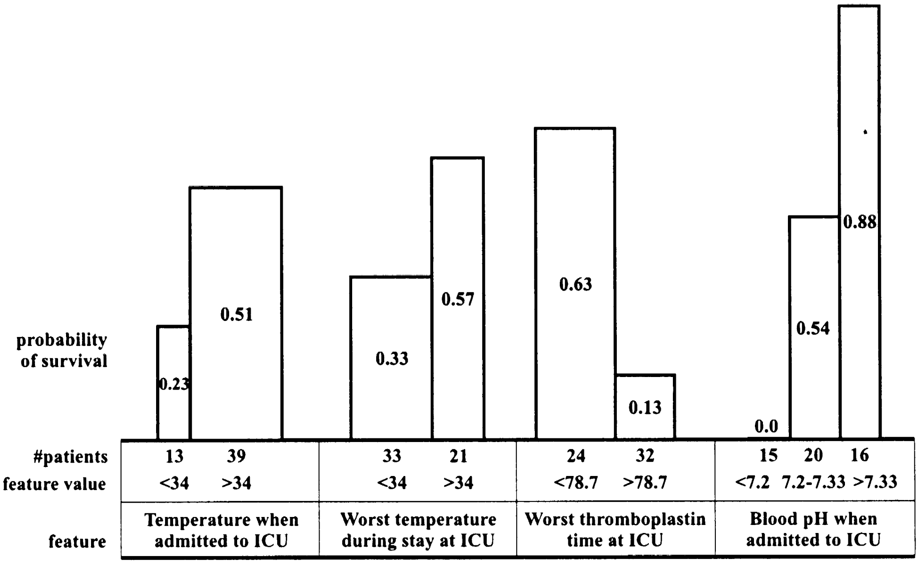

Fig. 2. Histograms that show distribution of patients and probabilities of their outcomes for different values of the body temperature when admitted to ICU, minimal body temperature while at ICU, the worst pH value at ICU, and the worst partial active thromboplastin time. The width of the column corresponds to the number of patients and its height to probability of survival.

value (PH_ICU), which reflects many important aspects of the injury (respiratory and cardiovascular distress, blood loss and cellular damage). As the lower value indicates severe damages to patient's vital systems, patients with pH level below 7.20 are not expected to survive. A normal value of pH (> 7.33) predicts probable survival of the patient. The outcome for the patients with the pH values between 7.20 and 7.33 is predicted from the worst partial active thromboplastin time value (APPT_WORST), which assigns a greater probability of survival to the patients with normal blood coagulation. Notice that the resulting tree incorporates only two of the nine predictive features from the data set used. We can, however, speculate that retrospective data that would include a higher number of patients would enable us to induce a larger, yet reasonable classification trees.

Naive Bayes classifier can be graphically represented in a device called a nomogram [9]. The nomogram (Fig. 3) shows the impact of individual features on death (upper labels on feature lines) and survival (lower labels). The values right of zero favor death/survival and the values on the left speak against it. For example, PH_ICU levels above 7.33 and between 7.20 and 7.33 are to the left of zero and speak against the patient's death, while values below 7.20 are indicators for death.

Nomogram can be used to compute the probabilities of outcomes. First, the impact factors for feature values must be summed, once for death and once for survival, using the scale above (below) the table. The sums are then converted into probability estimation using the lookup graph at the bottom of the nomogram and, finally, normalized to sum of 1. Features, which were not measured, can be simply ignored during computation. For example, a patient (APPT_ WORST = 85, BLEEDING_T = NO, PH_

ICU = 7.25) has the sum $0.3 - 0.2 - 0.2 = -0.1$ for death and $-1.0 + 0.3 + 0.3 = -0.4$ for survival. Approximation by the lookup table (scale at the bottom of a nomogram) gives 60% for death and about 22% for survival, which, when normalized to 100%, gives the final probabilities of 73% for death and 27% for survival.

The nomogram also points out some specifics about the domain we are modeling. The features whose values are most dispersed through the score line are the ones that are most predictive, i.e. influence the outcome most. In our case, the nomogram suggests the PH_ICU is the most important factor, followed by PT_ICU BE_ICU and APPT_WORST. Features BLEEDING_T and TYPE_OF_CL seem to have much smaller impact on the outcome. Interestingly, this is in accordance with classification tree from Fig. 1, which places the most relevant feature PH_ICU on the top, and additionally uses APPT_WORST.

5. Conclusion

This paper reports on a study, which has attempted to construct outcome prediction models from retrospective data of severe trauma patients. The study should be regarded as pilot since it only includes 68 patients. Despite having such small data set, the following conclusions can be drawn.

- A rather small subset of features from trauma patient's database seems sufficient for modeling.
- Given a proper selection of features, prognostic models for the outcome for severe trauma patients are plausible.

Furthermore, we show that the feature mining and machine learning techniques used identified two factors, namely the pH value when admitted to ICU and the worst partial active thromboplastin time, to be of highest importance for prediction. This finding is in accordance with previously published theoretical results [13]. While in theory the patient's body temperature should be another

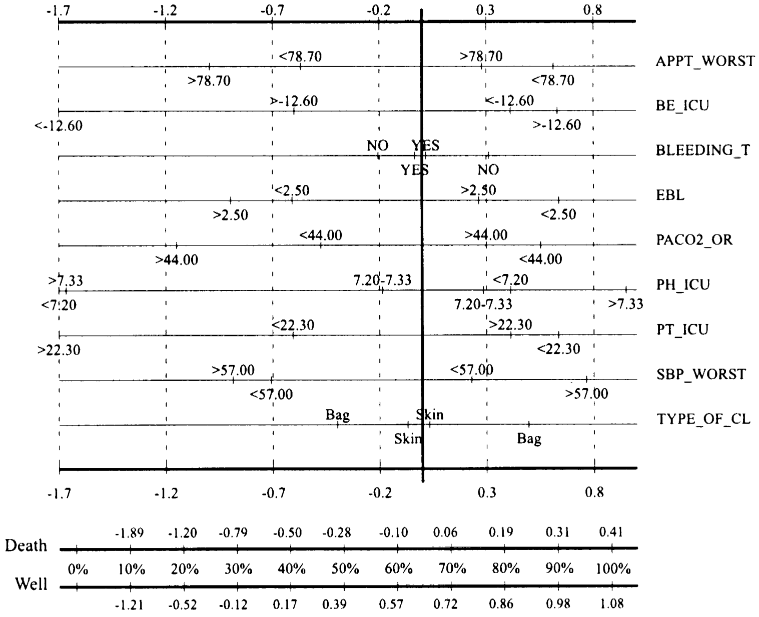

Fig. 3. A nomogram derived from naive Bayes classifier.

important predictive factor, other factors were found more predictive in our data set. Notice, however, that the work reported here was a feasibility study. Although the model showed good prediction accuracy, we must be cautious about applying such models in clinical practice because of the small number of patients. Further studies with larger numbers of patients are needed to confirm the results.

From methodological point of view, this study has found feature categorization and feature subset selection algorithms useful preprocessing techniques. Categorization of most relevant features was inspected and confirmed by the expert. Expert also found the feature rating by RELIEFF meaningful. This rating helped him to decide, which set of features should be used in the modeling data set. Both naive Bayes modeling and derivation of classification trees resulted in models of reasonable performance, with the models not being significantly different in performance.

The main result of the study reported is the observation that prognostic models can be built for prediction of outcomes for severe trauma patients. In future work, this finding needs to be verified in a study that would include a larger number of patients. The present data set includes many features, and if such comprehensive data collection poses problems as our present data set with many missing data suggests the outcome of this study may help trauma personnel to focus mostly on features that we have observed to be the most relevant for prediction.

References

[1] R. Belazzi, B. Zupan, Intelligent data analysis in medicine and pharmacology: a position statement, in: IDAMAP-98, Brighton, UK, 1999, pp. 1–4.

[2] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous valued attributes for classification learning, in: Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, Chambery, France, 1993, pp. 1022–1029.

[3] K. Kira, L.A. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proceedings of AAAI 92, San Jose, CA, 1992.

[4] I. Kononenko, Estimating attributes: analysis and extensions of relief, in: F. Bergadano, L. De Raedt (Eds.), Proceedings of the European Conference on Machine Learning (ECML-94), Springer, Berlin, 1994, pp. 171–182.

[5] I. Kononenko, On biases in estimating the multivalued attributes, in: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, 1995, pp. 1034–1040.

[6] I. Kononenko, I. Bratko, M. Kukar, Application of machine learning to medical diagnosis, in: Machine Learning and Data Mining: Methods and Applications, Wiley, Chichester, UK, 1998, pp. 389–408.

[7] I. Kononenko, B. Zupan, Attribute mining: evaluation, discretization, subset selection and constructive induction, in: 'From Machine Learning to Knowledge Discovery in Databases' Workshop Notes, at ICML-99, Bled, Slovenia, 1999.

[8] N. Lavrač, E. Keravnou, B. Zupan (Eds.), Intelligent Data Analysis in Medicine and Pharmacology, Kluwer Academic Publishers, Boston, 1997.

[9] J. Lubsen, J. Pool, E. van der Does, A practical device for the application of a diagnostic or prognostic function, Methods Inf. Med. 17 (1978) 127–129.

[10] P.J.F. Lucas, A. Abu-Hanna, Prognostic methods in medicine (editorial), Artif. Intell. Med. 15 (2) (1999) 105–119.

[11] D. Michie, D.J. Spiegelhalter, C.C. Taylor (Eds.), Machine Learning, Neural and Statistical Classification, Ellis Horwood, Chichester, UK, 1994.

[12] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.

[13] M.F. Rotonda, D.H. Zonies, The damage control sequence and underlying logic, Surg. Clin. North Am. 77 (1997) 761–777.

[14] B. Zupan, N. Lavrac, E. Keravnon, Data mining techniques and applications in medicine (editorial), Artif. Intel. Med. 16 (1999) 1–2.