© Schattauer GmbH (2001)

B. Zupan[1, 2, 3], J. Demšar[1], D. Smrke[4], K. Božikov[4], V. Stankovski[4], I. Bratko[1, 2], J. R. Beck[3]

Predicting Patient's Long-Term Clinical Status after Hip Arthroplasty Using Hierarchical Decision Modelling and Data Mining

[1]Faculty of Computer and Information Sciences, University of Ljubljana and
[2]J. Stefan Institute, Ljubljana, Slovenia,
[3]Office of Information Technology, Baylor College of Medicine, Houston, TX, USA,
[4]Department of Traumatology, University Clinical Center, Ljubljana, Slovenia

Abstract: Construction of a prognostic model is presented for the long-term outcome after femoral neck fracture treatment with implantation of hip endoprosthesis. While the model is induced from the follow-up data, we show that the use of additional expert knowledge is absolutely crucial to obtain good predictive accuracy. A schema is proposed where domain knowledge is encoded as a hierarchical decision model of which only a part is induced from the data while the rest is specified by the expert. Although applied to hip endoprosthesis domain, the proposed schema is general and can be used for the construction of other prognostic models where both follow-up data and human expertise is available.

Keywords: Harris Hip Score, Hip Arthroplasty, Prognostic Models, Data Mining, Hierarchical Decision Models

1. Introduction

The incidence of femoral neck fractures has doubled in the last 30 years along with an increase of the population over 65 years of age. An acute medial femoral neck fracture is most often treated with implantation of bipolar partial or total hip endoprosthesis [9, 12, 13]. The aim of treatment is to enable the patient to walk soon after surgery on a stable and painless hip, and to avoid avascular necrosis or nonunion, or both. Endoprosthesis as a primary treatment of the fracture can ensure lower number of complications after the operation, a shorter rehabilitation period, lower number of revision arthroplasties and can ultimately lead to a higher quality of life of the elderly patient.

Various factors like, for example, patient's health status prior to operation, timing, possible complications, and quality of rehabilitation may influence the final outcome after a successful implantation procedure. To systematically assess the significance of these factors and determine the way they relate to the patient's long-term clinical status, a study that included 112 patients admitted and operated at the Department of Traumatology of University Clinical Center in Ljubljana from January 1988 to December 1996 was conducted. For each patient, a number of features at the time of operation or immediately after were recorded. Patient's long-term clinical status was assessed at least 18 months after the operation. It has been shown [16] that the effects of the operation of implantation of hip endoprosthesis stabilize one year after the operation. We have used this observation in our study and required even greater minimum follow-up time. The importance of collected factors was studied through a construction of a prognostic model that would, immediately after the operation, be able to predict the patient's long-term status. The main questions were (1) which factors are most related to the outcome and, if properly selected, (2) whether they contain enough information to be included in a model of reasonable predictive performance. Although this article focuses on construction and selection of factors that are relevant to outcome prediction, the model that resulted from the study may be used in clinical practice for early identification of critical cases and, prior to operation, to study different rehabilitation scenarios.

Recently, various methods from the research areas of machine learning and data mining have been developed that may use existing patient data to construct prognostic models. Different studies have shown, however, that using additional explicitly encoded expert knowledge can significantly enhance the performance of induced models [14, 24]. For the hip arthroplasty domain, our expert physician (co-author of this paper D. S.) expressed her domain knowledge by grouping functionally-related features, further representing each group with a new feature at a higher level of abstraction, and organizing original and new features into a feature hierarchy. We have borrowed this background-knowledge elicitation method from the field of decision support systems, and specifically from hierarchical multi-attribute decision systems [2, 3, 20]. This approach assumes that the expert can both develop a feature hierarchy and define all relations between features. Although we have found that it was relatively easy for a physician to relate the original and new features

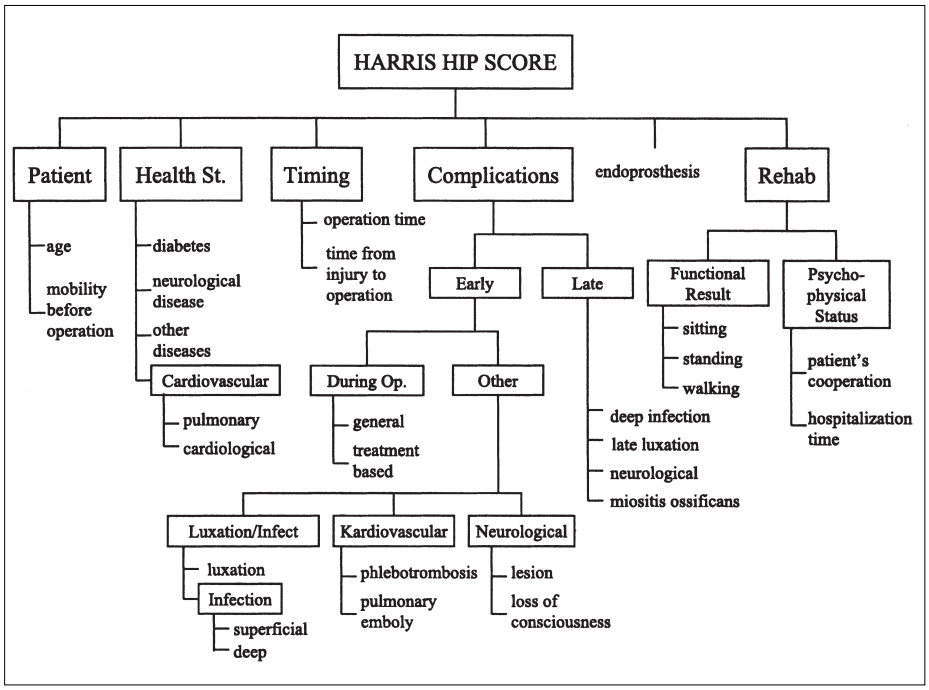


Fig. 1 Feature hierachy.

by providing a corresponding mapping function, the use of a machine learning algorithm was required to mine the data and relate the selected higher-level features to the patient's long-term status. The methodology and the main contribution of the work presented is thus a proposed combination of hierarchical decision modeling and machine learning to construct prognostic models of improved performance.

The article is organized as follows. We first present the characteristics of the data that was used to induce the prognostic model. The feature hierarchy as defined by the physician is discussed next. Using the data and selected features from the feature hierarchy, we study the utility of two different machine learning algorithms to construct a prognostic model. Experimental evaluation shows that a combination of data mining and hierarchical decision modeling may significantly improve the performance of a prognostic model and lead to a model that can be used in clinical practice.

2. Data Set

A data set of 112 patients admitted and operated at the Department of Traumatology of University Clinical Center in Ljubljana from January 1988 to December 1996 was collected. For each patient, a study recorded a set of features at the time or immediately after operation. We will refer to them as *recorded* or *measured features*. For the study reported in this article, 28 features that may potentially influence the long-term clinical status of the patient were selected by physician. These mainly include patient's health status prior to operation, data about complications prior, during, and after the operation, type of implanted endoprosthesis, and results of early rehabilitation. Most of the features recorded were categorical (e.g., "pulmonary diseases" having possible values "no" and "yes"). For a few originally numeric features the values were categorized by the physician by expressing meaningful cut-off points (e.g., for "walking" the cut-off point is 9 days, and the feature has two categorical values of "less or equal to 9 days" and "over 9 days"). Categorization allows the introduction of new qualitative labels that are used in physicians' daily practice or are known from the literature. The categorization was also required by the DEX tool used in defining the qualitative decision model (Section 3).

The long-term clinical status was assessed through a follow-up not earlier than 18 months after the operation and used a scoring system proposed by Harris [8, 10]. The *Harris hip score* gives an overall assessment of the patient's condition and is evaluated by a physician who considers patient's ability to walk and climb stairs, patient's overall mobility and activity, presence of pain, function of the hip, quality of life, and absence of deformations in general with respect to the injured hip. It uses a point scale from 0 to 100 – the higher the score, the better the long-term outcome after hip arthroplasty. For the purposes of this study, the Harris hip score was categorized to "bad" (values below or equal to 70), "good" (values above 70 but lower than 90) and "excellent" (values above 90). The follow-up study included 38.4% patients with bad, 30.4% with good, and 31.2% with excellent Harris hip score. The cut-off points used were suggested by the expert physician according to her own opinion and related literature [8, 10], as they introduce three particular and meaningful outcome categories. Namely, the patients with "bad" Harris hip score are constantly experiencing problems with their hips (pain, inability for a longer walk, restricted function of the hip); patients with "good" Harris hip score have some problems with the hip occasionally; "excellent" Harris hip score means that the patient has almost no problems at all with the hip, may engage in recreational sports, etc.

3. Hierarchical Decision Model

To express additional (background) knowledge about the hip arthroplasty domain, the physician was asked to organize recorded features into meaningful groups. Not surprisingly, this was a relatively easy task for the physician, since these groups were already partially identified in the questionnaire that was used to collect patient's data. As there were many groups that include only a few features, the physician was further asked to further organize some of the groups. As the identified groups may represent new intermediate features, the procedure described above resulted in a hierarchical classification of features (called a *feature hierarchy*) that is shown in Fig. 1. To distinguish the originally recorded features at the leaves of the feature hierarchy from the names of intermediate features, the latter ones are enclosed in boxes.

For example, features that recorded the time of first sitting, standing and walking of the patient after the operation were grouped within "Functional Result". So "Functional Result" depends on sitting, standing, and walking. Similarly, as shown in the hierarchy from Fig. 1, "Psychophysical status" depends on patient's cooperation and hospitalization time. The physician further identified that "Functional Result" and "Psychophysical Status" both define the length of "Rehabilitation", thus introducing another intermediate feature.

Besides naming each feature group, intermediate features were also assigned a set of possible values. For example, "Functional Result" may be "good" or "bad", the length of "Rehabilitation" may be "short", "medium", or "long", etc.

For most of the newly introduced intermediate features, the physician was able to identify their functional relation with their immediate successors in the feature hierarchy. The relations were expressed as pointwise *utility functions.* For example, Table 1 shows such utility functions for intermediate features that encode complicationser during operation ("During op.") and early complications ("Early"). For instance, the second row in the left table in Table 1 says that when there are no general complications ("general = no") but there are treatment-based complications ("treatment based = yes"), it is considered that there are complications during operation ("During op. = yes"). The utility functions were completely specified, i. e., the physician assigned a value to an intermediate feature for every combination of values of the preceding features in the hierarchy.

Notice that utility functions can not be induced from the data by standard machine learning techniques (like decision tree induction, rule induction, etc.). This is simply because such algorithms require a class value to be given in the examples. Our intermediate features are not contained in the data and the only knowledge about them is their name, set of values they use and their inclusion in a feature tree. There exists a technique called function decomposition [22] that may be used for this task. However, for the problem reported in this article, we preferred to use utility functions defined by the expert physician.

Table 1a Utility functions for complications during the operation.

general	treatment based	during op.
yes	yes	yes
no	yes	yes
yes	no	yes
no	no	no

Table 1b Utility functions for early complications.

during op.	other	early
yes	many	many
no	many	few
yes	few	many
no	few	few
yes	none	many
no	none	none

The main reason for this was that she found such encoding of her domain knowledge very convenient and was confident that her utility functions were meaningful and valid. Moreover, in the analysis of medical data it is often preferred to use all available domain knowledge [1]. The comparison of "manually" developed utility functions and those that may be discovered through function decomposition was beyond the scope of this paper, and belongs to future work.

Using feature hierarchy and its corresponding utility functions, and given the values of the measured features for a specific patient, the values for intermediate features may be determined. For example, if the patient had no general and treatment-based complications, this means there were no complications during the operation (Table 1). If we similarly find that the same patient had a few other complications, we can conclude that she had a few early complications (Table 1).

Utility functions were elicited from the physician for all intermediate features except for the outcome feature, that is, for the Harris hip score. It was only pointed out by the physician that the Harris hip score may depend on the characteristics of the patient ("Patient"), her health status ("Health St."), timing of the operation ("Timing"), complications prior, during and after the operation ("Complications"), type of endoprosthesis ("Endoprosthesis"), and requirements for rehabilitation ("Rehab"). Thus, the relation between the measured and intermediate features and the outcome needed to be determined from patient's data, as shown in the next Section.

It took about one day of the physician's time to define the feature hierarchy, and about one day to define the utility functions. For both tasks, we have used a decision-support system shell DEX [2], that supports graphical construction of the hierarchy and, more importantly, actively supports the construction of utility functions by performing consistency checks, suggesting the values of the intermediate features for the combinations not yet specified by the expert, and being able to summarize the utility functions graphically or textually [2, 3]. We further found DEX useful for the actual evaluation of patient's data. Namely, given a data set of 28 original features for 112 patients, DEX augmented this set with the values for all 16 intermediate features introduced in the feature hierarchy and defined through the use of physician-specified utility functions. The resulting data set thus consisted of 44 features and a corresponding outcome (categorized Harris hip score).

4. Data Mining and Induction of Prognostic Model

Data mining was used to determine which features are most relevant and should be used in constructing a prognostic model of acceptable performance. First, the information gain was assessed for each feature. *Information gain* was proposed by Quinlan [19] and in essence measures how well can one distinguish between different values of the outcome by knowing only the value of a single feature. The higher the information gain, the more relevant the feature is to determine the outcome.

All 44 features, that is, all the original and intermediate features from the feature hierarchy were assessed. Ten most relevant features are presented in Table 2. It is interesting that among the highest-ranked features most of them are intermediate ones. Although this result was expected, it also confirms the potential high value of intermediate

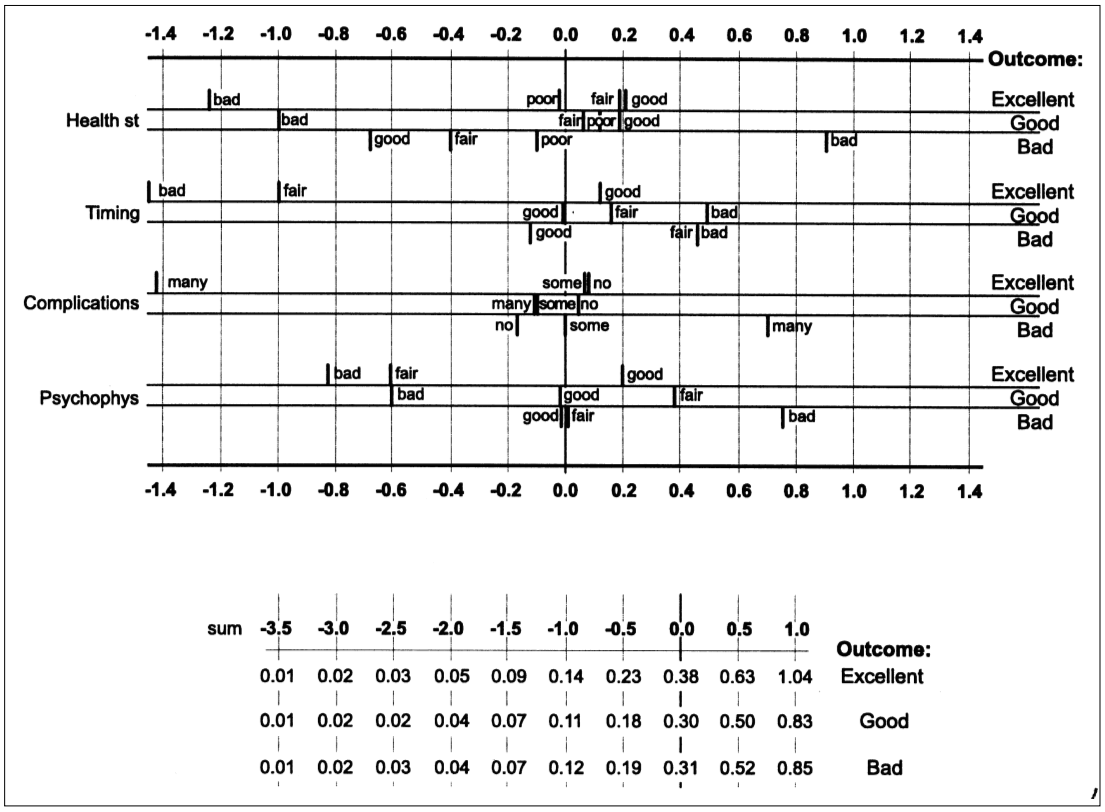


Fig. 2 Nomogram for naive Bayes' prognostic model.

features and may indicate for correctness of utility functions that express them. Namely, since each intermediate feature in essence combines at least two original features, it should – if properly encoded and constructed – give more information about the outcome than each of the features it combines. Let us illustrate this point with the following example. Among the best-rated features are measured features "time from injury to operation" and "hospitalization time". But these two features are constituents of intermediate features "Timing" and "Psychophysical Status", respectively, which are both rated higher. Similarly, "Cardiovascular" is rated higher than "pulmonary", but lower than "Health St.", and "Complications" is rated higher than "Neurological".

An exception of the above is "sitting", the fourth measured feature included in the list of ten most relevant. After the inspection of the database, we found, however, that the faster that the patients are able to sit, the lower is their outcome Harris hip score. This is in contradiction to the common expert knowledge for hip arthroplasty domain, and identifies the potential problem with undersampling of the problem space for this particular feature. Since this common expert knowledge was encoded in a utility function for "Functional Result", "sitting" in a way spoiled the value for "Functional Result" when combined with "standing" and "walking," and thus "Functional Result" was rated rather low.

We next needed to select the features for our target prognostic model. From the ten best-rated features from Fig. 1, we have decided to include only the features of the highest level of aggregation. Using this rule, only the features "Health St.", "Timing", "Complications", and "Psychophysiological Status" were used. Features "pulmonary" and "Cardiovascular" were left out because of the higher-level feature "Health St." We excluded "Neurological" complications since they are constituents of "Complications". For a similar reason we do not use "time from injury to operation" and "hospitalization time". Since, as stated above, we have found inconsistencies with feature "sitting", we do not include it in the final model.

Besides relying on the above evaluation of feature relevance based on the information gain, the decision to include the features at the highest level of aggregation stems from the work in research areas of hierarchical decision models [2, 20] and structured induction [21]. There, the outcome is always directly related to the features at the highest level. For the particular problem of Harris hip score prediction, we further experimentally show that such feature selection may result in the most accurate prognostic models (Section 5).

Two different standard machine learning techniques were considered for derivation of the prognostic model from data. For *decision tree induction*, we have used our own implementation of the ID3 recursive partitioning algorithm [19]. The particular decision tree induction algorithm included pre- and post-pruning. In pre-pruning, the recursive partitioning was stopped if it would build leaves consisting of less than two data records. For post-pruning, the minimal-error pruning algorithm [18] using the *m*-probability estimate [4] with $m = 10.0$ was used.

As an alternative to decision trees, the naive Bayes' machine learning method [11] was used to induce a prognostic model. Assuming the independence between features, the probability that a

patient described with the value vector $V = (v_1 \ldots v_n)$ of features has an outcome $o \in O$, where O is a set of possible outcome values, can be estimated by Bayes' formula

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

where $P(o)$ is the prior probability of an outcome o and $P(o|v_i)$ is the conditional probability of an outcome o given i-th predictor variable has the value v_i; both are estimated from the data set of patients. This formula can be derived from the more common form $P(o|V) = [P(o)/P(V)] \prod_i P(v_i|o)$ by reapplying the Bayes rule $P(v_i|o) = P(o|v_i) P(v_i)/P(o)$, assuming the complete independence of attributes.

An experimental comparison of the two machine learning algorithms, as presented in the next section, showed that naive Bayes' models were more accurate. For our problem, the model developed with the naive Bayes' algorithm is thus the model of choice and we describe it here in more detail. Derived naive Bayes' prognostic model can be graphically presented as a nomogram [15, 23]. The nomogram (Fig. 2) shows the impact of individual features on the probability of each of the three outcome values. The positions of feature values are computed as logarithms of their respective terms in the Bayes' formula; e.g., the left-most value of the upper-most axis of Timing shows the impact of bad timing on excellent outcome, and its position is computed as ln(*P*[excellent|Health St = bad]/*P*[excellent]). Thus, the values right of zero favor the specific outcome and the values on the left speak against it.

The nomogram can be used to manually compute the probabilities of specific outcomes for a particular patient. For example, let us take a patient with good health status, fair timing, no complications and good psychophysiological status. As the nomogram shows logarithms of conditional probabilities, multiplying probabilities in the original Bayes' formula corresponds to summing up the values from the nomogram. Impacts of individual factors on the probability of excellent outcome are plotted on the upper axis (labeled "Excellent"). For our case, summing up the impacts of the given values, that is +0.2 for health status, –1.0 for timing, +0.1 for complications and +0.2 for psychophysiological status, gives a total of –0.5. The negative sum already suggests that the probability for the high value of the Harris hip score for this patient is lower than the average and that the major factor causing it is fair timing. The scale below the nomogram is provided for converting the sums into probabilities of outcomes. For our sum of –0.5, the probability for outcome "Excellent" is 23%.

We proceed by summing the impacts for outcomes "good" and "bad", plotted on the middle and lower axis. For outcome "good", we have 0.18 + 0.16 + 0.05 – 0.02 = 0.37, which translates to probability of about 40%, and for outcome "bad" we have –0.68 + 0.45 – 0.17 – 0.02 = -0.42, giving a probability of about 20%. Note that the Bayes' formula is an approximation, so the computed probabilities for all three outcomes do not necessarily sum to 100% (in our case, their sum is 83%).

Additional information that can be extracted from the nomogram is the importance of individual features and individual values. The "Psychophysiological Status" seems to be less important, as the impacts of its possible values are limited to a smaller region than the values of the other three features. More interestingly, there are no feature values, which can significantly improve the Harris hip score as the values with positive influence are at most at +0.2, which is low in comparison with the values that decrease the chances for a successful implantation and can be more than –1.4. The attributes with a significant negative influence are bad health status, bad or fair timing, many complications and bad or fair psychophysiological status. An undesired value of any feature will always outweigh positive values of others. In other words, a complication in some respect cannot be compensated by favorable values in other respects.

Table 2 Ten most relevant features to determine the long-term status of the patient.

Feature name	Information gain
Health St.	0.139
Cardiovascular	0.113
Pulmonary	0.095
Psychophysical Status	0.065
Timing	0.060
Time from injury to operation	0.052
Hospitalization time	0.050
Complications	0.048
Neurological	0.047
Sitting	0.046

5. Experiments and Discussion

In this section we compare the performance of naive Bayes' models and decision tree models in our domain. We show that naive Bayes' models may perform significantly better than induced decision trees.

To evaluate both machine learning approaches and the particular selection of features, a standard technique of stratified 10-fold cross-validation was used [17]. This divides the patient data set to 10 sets of approximately equal size and equal distribution of outcomes.

Table 3 Classification accuracy of different prediction models.

Features used	Naive Bayes'	Decision Tree
All measured features	44.6	42.9
10 most informative features	55.4	51.8
4 features selected from best 10	56.3	51.8

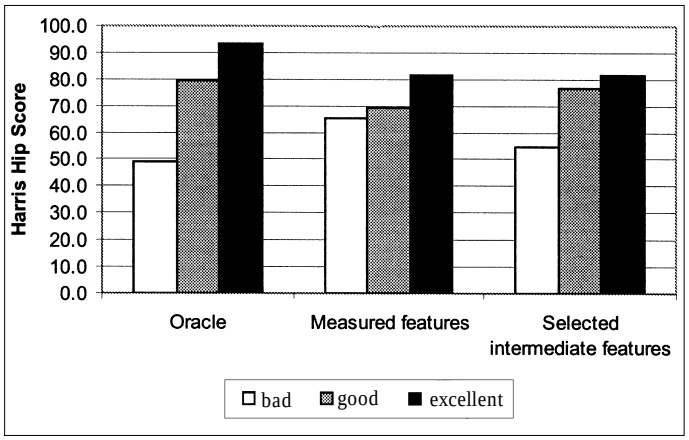


Fig. 3 Average Harris hip scores for the predicted outcome class. Oracle ("true" class for each patient) is compared with class predicted by naive Bayes when using either 28 original features or only a set of four selected intermediate features.

In each experiment, a single set is used to measure the classification accuracy of the model that has been developed from the remaining nine sets. *Classification accuracy* is expressed in percent of patients in the test set that were classified correctly. Note that since our database has a majority of 38.4% patients with a bad Harris hip score, it is expected that a reasonable classifier should significantly exceed this lower-bound for classification accuracy.

We have tried to learn the prognostic the model from 28 original features, from the best ten of the measured and intermediate features (Table 2), and from a selection of intermediate features taken from the list of best ten features as advocated in the previous section. The results are presented in Table 3.

Notice that all induced models perform better than the default classifier that would always classify to the majority class and thus obtain classification accuracy of 38.4%. For any feature selection used, the naive Bayes' learner always outscored decision tree induction, though when the significance was measured with the McNemar test [5, 7] the difference was not significant (p = 0.18 when four selected features are used). The highest scoring algorithm is naive Bayes when learning from the set of four selected features. The performance is significantly better (p <0.01) when compared to default classifier, and also significantly better (p <0.05) than the naive Bayes' when using only the measured features. This result clearly points out the value of additional domain knowledge in the form of feature hierarchy.

The superior performance of the naive Bayes' method when compared to induction of decision trees was not unexpected. As was shown in the study by Kononenko et al. [11], naive Bayes often performs better than other more sophisticated machine learning algorithms when learning in medical domains. As early as in the preparation of the study, medical experts may be inclined to choose relatively independent features or at least features that do not have complex interactions. For such data sets, even when the features are not really independent, it has been shown that naive Bayes' classifier performs well [6]. Furthermore, the features carefully crafted and selected by physicians and included in medical data sets, are usually all more or less relevant to the outcome. In this respect, the particular disadvantage of decision tree is that its branches typically consider only a small subset of features, while the naive Bayes' model always takes into account all features that were used for learning.

Figure 3 shows the numerical value of the Harris hip score averaged across patients that were classified to the same outcome class. This analysis shows that the model induced from measured features poorly differentiates patients with "bad" and "good" outcome, while the model induced from intermediate features performs better in this respect. The major reason for the success of the latter model can thus be contributed to the improved classification of patients with "bad" outcomes. Notice also that the model with selected intermediate features predicts "good" outcome to the patients that indeed have the higher Harris hip score (compared to the model with all measured features, the average Harris hip score of such patients is raised from 70 to 78). We can further see that for this model the average scores for "bad" and "good" patients are relatively close to the expected ones (54 instead of 49 for "bad" and 78 instead of 80 for "good"). Its weak point is the prediction to "excellent" outcome, where our selected model does not improve upon the model that uses the measured features.

6. Conclusion

In general, medicine may be considered as a data and knowledge-rich environment [1]. When constructing prognostic models for specific medical domains, it may thus be useful to both learn a model from the data while additionally using an explicitly coded expert knowledge. This was also the idea of the approach we presented in this article. The task was to construct a prognostic model for a long-term outcome after acute femoral neck fracture treatment with implantation of hip endoprosthesis. The expert knowledge was encoded as a hierarchy of patient's features recorded before or immediately after the operation. Additional knowledge was expressed with utility functions that relate measured features with new intermediate features used in the hierarchy. Selected intermediate features were then used to construct the model from patient data using naive Bayes' approach.

The resulting prognostic model thus consists of feature hierarchy and corresponding utility functions defined by the expert, and a data-induced utility function that models the outcome using four intermediate features. We show that an immediate advantage of using expert-defined background knowledge was in our case significantly improved performance of the prognostic model. We also show that without the background knowledge, the performance of the prognostic model would only be marginally better than a default classifier that would classify all patients to the same majority class.

Although applied to the hip endoprosthesis domain, the proposed schema that combines manual construction of a hierarchical decision model and its partial induction from the follow-up data is

general and can be used for construction of other prognostic models where both follow-up data and human expertise is available. The experience from the field of decision support systems also indicates that building a hierarchical decision model usually requires up to five days of expert's involvement. In our case it took two days. Additionally, there are also techniques available to support the induction of feature hierarchies from data [22] that can be used to further support the elicitation of the background knowledge.

The presented article focused on the construction of the prognostic model for outcome of hip arthroplasty. While our measure of success was based on estimated accuracy of the model, studies are under way that will assess the utility of PC-based and PalmPilot-based implementations of a prognostic model in everyday clinical practice.

REFERENCES

1. Bellazzi R, Zupan B. Intelligent data analysis in medicine and pharmacology: A position statement. In: IDAMAP-98. Brighton: UK 1988: 1-4.
2. Bohanec M, Rajkovič V. DEX: An expert system shell for decision support. Sistemica 1990; 1 (1): 145-57.
3. Bohanec M, Zupan B, Rajkovič. Hierarchical multi-attribute decision models and their application in health care. In: Medical Informatics Europe '99. IOS Press 1999; 670-5.
4. Cestnik B, Bratko I. On estimating probabilities in tree pruning. In: Proc European Working Session on Learning EWSL-91. Berlin: Springer-Verlag 1991; 1: 138-50.
5. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computa 1989; 10: 1895-924.
6. Domingos P, Pazzani M. Beyond independence: conditions for the optimality of the simple Bayesian classifier. In: Proceedings of the Thirteenth International Conference on Machine Learning. Bari: Morgan Kaufmann 1996; 105-12.
7. Everitt BS. The Analysis of contingency tables. London: Chapman and Hall 1977.
8. Harris WH. Traumatic arthritis of the hip after dislocation and acetabular fractures: Treatment by mold arthroplasty: end result study using a new method of result evaluation. J Bone Joint Surg 1969; 51-A: 737-55.
9. James SE, Gallannaug SC. Bi-articular hemiarthroplasty of the hip: a 7-year follow-up. Injury 1991; 22 (5): 391-3.
10. Johnston RC, Fitzgerald RH, Poss R, Muller ME, Sledge CB. Clinical and radiographic evaluation of total hip replacement. A standard system of terminology for reporting results. J Bone Joint Surg 1990; 72 (2): 161-8.
11. Kononenko I, Bratko I, Kukar M. Application of machine learning to medical diagnosis. In: Machine Learning and Data Mining: Methods and Applications. Chichester: John Wiley & Sons 1998; 389-408.
12. Koval KJ, Skovron ML, Aharonoff GB, Zukerman JD. Predictions of functional recovery after hip fracture in the elderly. Clin Orthop 1998; 348: 22-8.
13. Koval KJ, Zukerman JD. Hip fractures are an increasingly important public health problem (editorial). Clin Orthop 1998; 348: 2.
14. Lavrač N, Džeroski S, Pirnat V, Križman V. The utility of background knowledge in learning medical diagnostic rules. Appli Artifi Intelli 1993; 7: 273-93.
15. Lubsen J, Pool J, van der Does E. A practical device for the application of a diagnostic or prognostic function. Method Inform Med 1978; 17: 127-9.
16. Lyons AR. Clinical outcomes and treatment of hip fractures. Am J Med 1997; 103 (2A): 51S-63.
17. Michie D, Spiegelhalter DJ, Taylor CC, ed. Machine learning, neural and statistical classification. London: Ellis Horwood 1994.
18. Niblett T, Bratko I. Learning decision rules in noisy domains. In: Expert Systems 86. Cambridge University Press 1986; 15-8. (Proc. EWSL 1986, Brighton).
19. Quinlan R. Induction of decision trees. Machine Learning 1986; 1 (1): 81-106.
20. Saaty TL. Multicriteria decision making: The analytic hierarchy process. Pittsburgh: RWS Publications 1992.
21. Shapiro AD. Structured induction in expert systems. Wokingham, UK: Turing Institute Press in association with Addison-Wesley Publishing Company 1987.
22. Zupan B, Bohanec M, Demšar J, Bratko I. Learning by discovering concept hierarchies. Artifi Intelli 1999; 109 (1-2): 211-42.
23. Zupan B, Demšar J, Kattan MW, Beck JR, Bratko I. Machine learning for survival analysis: A case study on recurrence of prostate cancer. In: AIMDM-99. Horn W, et al., ed. Heidelberg: Springer 1999: 346-55.
24. Zupan B, Džeroski S. Acquiring and validating background knowledge for machine learning using function decomposition. Artifici Intelli Med 1998; 14 (1-2): 101-17.

Address of the authors:
Dr. Blaz Zupan,
Faculty of Computer and Information Science,
University of Ljubljana,
Trzaska 25,
SI-1000 Ljubljana,
Slovenia
E-mail: blaz.zupan@fri.uni-lj.si