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

B Zupan [1,2,3], J Demsar [1], D Smrke [4], K Bozikov [4], V Stankovski [4], I Bratko [1,2], JR Beck [3]

[1] Faculty of Computer and Information Sciences, University of Ljubljana, Slovenia

[2] J. Stefan Institute, Ljubljana, Slovenia

[3] Office of Information Technology, Baylor College of Medicine, Houston, TX

[4] Department of Traumatology, University Clinical Center, Ljubljana, Slovenia

Abstract

The article presents a construction of a prognostic model 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. The article proposes a schema 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 construction of other prognostic models where both follow-up data and human expertise is available.

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 [11, 10, 7]. 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 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. The importance of collected factors was studied through a construction of a prognostic model, that would, just after the operation, be able to predict 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 the 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, 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 showed, however, that using additional explicitly encoded expert knowledge can significantly enhance the performance of the induced models [12, 19]. For hip arthroplasty domain, the physician 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 factor hierarchy. We have borrowed this background-knowledge elicitation method from the field of decision support systems, and specifically from hierarchical multi-attribute decision systems [16, 2, 3]. The latter assume that the expert can develop both feature hierarchy while manually encoding all relations between features. Although we have found that it was relatively easy for a physician to relate original and new factors by providing a corresponding mapping function, the use of machine learning algorithm was required to mine the data and relate the selected higher-abstraction 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 modelling 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 physician is discussed next. Using the data and selected features from the feature hierarchy, we use a naive Bayesian machine learning algorithm to construct a prognostic model. Evaluation shows that a combination of data mining and hierarchical decision modeling may significantly improve the performance of prognostic model and lead to a model that can be used in clinical practice.

2 The Data

A dataset of 112 patients admitted and operated at Department of Traumatology of University Clinical Center in Ljubljana from January 1988 to December 1996 was collected. For each patient, a study recorded over 100 features at the time or immediately after operation.

We will refer to them as to *recorded* or *measured features.* Of these, 28 features that may have an influence to 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 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").

The long-term clinical status was assessed through a follow-up at least 18 months after the operation and used a scoring system proposed by Harris [6, 8]. *Harris hip score* gives an overall assessment of 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 and 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, 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.

3 Hierarchical Decision Model

To express additional (background) knowledge about the hip arthroplasty domain, physician was asked to organize recorded features into meaningful subgroups. Not surprisingly, this was a relatively easy task for the physician, since these groups were already partially identified in a questionnaire that was used to collect patient's data. As there were rather many groups that include only a few features, 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 Figure 1. To distinguish originally recorded features at the leaves of the feature hierarchy from the names of intermediate features, the last ones are outlined.

For example, features that recorded the time of first sitting, standing and walking of the patient after the operation were grouped within "Functional Result". It can be said that "Functional Result" depends on sitting, standing, and walking. Similarly, as shown in the hierarchy from Figure 1, "Psychophysical status" depends on patient's cooperation and hospitalization time. 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

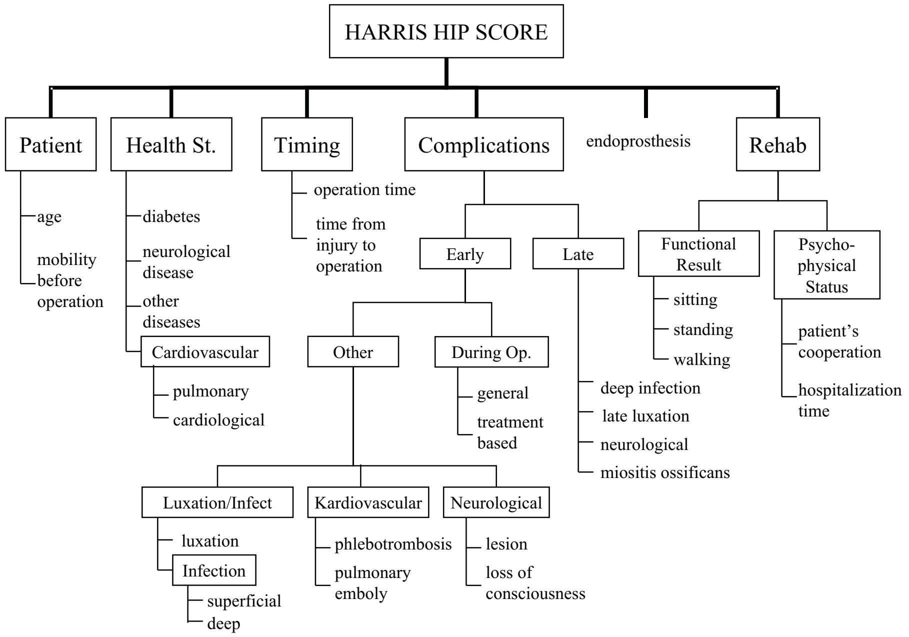

Figure 1: Feature hierarchy

general	treatment based	During op.
yes	yes	yes
no	yes	yes
yes	no	yes
no	no	no

During op.	Other	Early
yes	many	many
no	many	few
yes	few	many
no	few	few
yes	none	many
no	none	none

Figure 2: Utility functions for complications during the operation (left) and early complications (right).

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, 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, Figure 2 shows such utility functions for intermediate features that encode complications during operation ("During op.") and early complications ("Early"). For instance, second row in the left table in Figure 2 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., physician assigned a value to an intermediate feature for every combination of values of preceding features in a hierarchy.

Using feature hierarchy and its corresponding utility functions, and given the values of measured features for specific patient, the value 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 (Figure 2, Table on the left). If we similarly find that the same patient had few other complications, we can conclude that she had a few early complications (Figure 2, Table on the right).

Utility functions were elicited from the physician for all intermediate features but for the outcome feature which encodes Harris hip score. It was however pointed out that the long-term outcome 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 measured and intermediate features on one side and the outcome on the other side needed to be determined from patient's data, as shown in the next Section.

It took about a day for the physician to define a feature hierarchy, and about a day to define 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 dataset 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 dataset 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 answer two questions: (1) which features are most relevant when trying to determine patient's long-term status, and (2) which features should be used to construct a prognostic model of acceptable performance. To answer the first question, the information gain was assessed for each feature. *Information gain* was proposed by Quinlan [15] 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 is the feature to determine the outcome.

All 44 features, i.e., original and intermediate features, were assessed. Ten most relevant features are presented in Table 1. 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 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 then “pulmonary”, but lower than “Health St.”.

Feature name	Information gain
Health St.	0.139
Cardiovascular	0.113
pulmonary	0.095
Psychophysical Status	0.065
Timing	0.060
time form injury to operation	0.052
hospitalization time	0.050
Complications	0.048
Neurological	0.047
sitting	0.046

Table 1: Ten most relevant features to determine the long-term status of the patient.

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 patient’s 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 Figure 1, we have decided to include only the features of the highest level of abstraction. 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-abstract feature “Health St.” We excluded “Neurological” complications since they are constituents of “Complications”. For a similar reason we do not use “time form 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.

The prognostic model was derived from the data using a naive Bayesian machine learning method [9]. Assuming the independence of features, the probability that a patient described with values of predictor features $V = (v_1 ... v_n)$ has an outcome $o \in O$, where O is a set of possible outcome values, can be estimated by Bayesian formula

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

where $P(o)$ is the apriori probability of an outcome o and $P(o|v_i)$ is the conditional probability of an outcome o if i-th predictor variable has the value v_i; both are estimated from the dataset of patients. Note that 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 Bayesian rule $P(v_i|o) = P(o|v_i)P(v_i)/P(o)$.

Derived naive Bayesian prognostic model can be graphically presented as a nomogram [18, 13]. The nomogram (Figure 3) 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 Bayesian formula; e.g., the left-most value of the upper-most axis shows the impact of bad timing on excellent outcome, and its position is computed as $\ln(P(excellent|HealthSt = bad)/P(excellent))$. Thus, the values right of zero favor the specific outcome and the values on the left speak against it.

Nomogram can be used to manually compute probabilities of specific outcomes for a particular patient. For example, let us compute the probability for the high value of Harris hip score given that a patient has good health status, fair timing, no complications and good psychophysiological status. Summing the impacts of the given values, that is $+0.21$ for health status, -1.0 for timing, $+0.09$ for complications and $+0.1$ for psychophysiological status, gives a total of -0.6. The negative sum already tells that the probability for the high value of 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 -0.6, the probability for an excellent outcome is around 20% (between 23% and 14%). Note that the Bayesian formula is an approximation, so the computed probabilities for all the three outcomes do not necessarily sum to 100%.

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 influences 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.

5 Evaluation

In the previous Section we have assumed that the reasoning behind the selection of features (choosing the best ranked higher-abstraction features) and utility of naive Bayesian machine

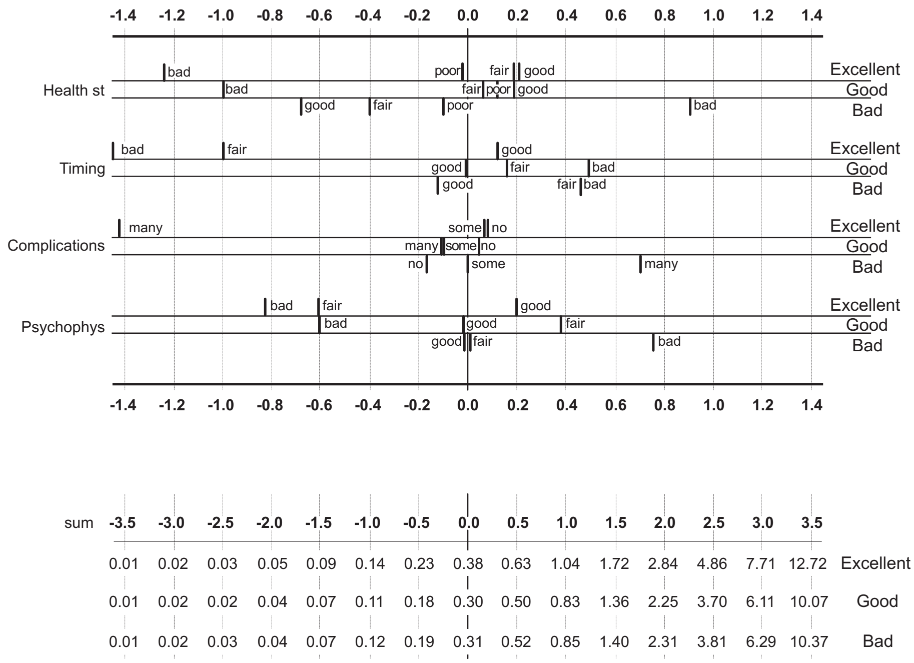

Figure 3: Nomogram for naive Bayesian prognostic model

learning leads to a plausible prognostic model. We here experimentally evaluate the two assumptions and assess the significance of our results.

To evaluate naive Bayesian method and the selection of features, a standard technique of stratified 10-fold cross-validation was used [14]. This divides the patient data set to 10 sets of approximately equal size and equal distribution of outcomes. 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 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 model from 28 original features, from best ten of measured and selected features (Table 1), 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 2. It is clear that naive Bayesian learner performs best when using the four features selected from the highest ranked features. The significance measured with McNemar test [5, 4] further shows that such classifier is significantly better ($p < 0.003$) than default classifier that always classifies to the majority class, and also significantly better ($p < 0.03$) than 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.

Figure 4 shows the numerical value of Harris hip score averaged across patients that were classified to the same outcome class. This analysis shows that the model induced from mea-

model	classification accuracy
majority class (default)	38.4
naive Bayes (measured features)	44.6
naive Bayes (best 10)	55.4
naive Bayes (selected from best 10)	56.3

Table 2: Classification accuracy of different prediction models.

sured 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 later model can thus be contributed to the improved classification of patients with "bad" outcomes.

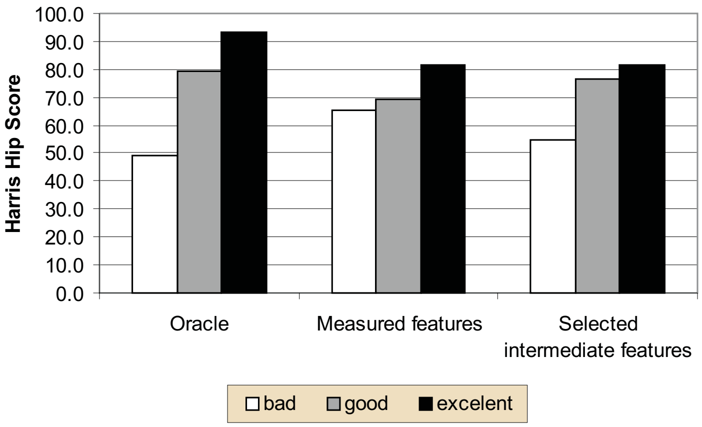

Figure 4: Average Harris hip scores for the predicted outcome class. Oracle ("true" class for each patient) is compared with class predicted by naive Bayes when using either 28 original features or only a set of four selected intermediate features.

6 Conclusion

In general, medicine may be considered as data and knowledge rich environment [1]. When constructing prognostic models for specific medical domain, it may thus be useful to both learn the model from the data while additionally using an explicitly coded expert knowledge. This was also the core idea of the approach we present in this article. The task was to construct 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 where then used to construct the model from patient data using naive Bayesian machine learning approach.

The resulting prognostic model thus consists of feature hierarchy and corresponding utility functions defined by 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 prognostic model. We also show that without the background knowledge, the performance of 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 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 can be cost-effective as it usually requires up to five days of expert's involvement (in our case only two days). Additionally, there are also techniques available to support the induction of feature hierarchies from data [17] 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 performance of the model, studies are under way that will assess the utility of PC-based and PalmPilot-based implementations of prognostic model in everyday clinical practice.

References

[1] R. Bellazzi and B. Zupan. Intelligent data analysis in medicine and pharmacology: A position statement. In *IDAMAP-98*, pages 1–4, Brighton, UK, 1988.

[2] M. Bohanec and V. Rajkovič. DEX: An expert system shell for decision support. *Sistemica*, 1(1):145–157, 1990.

[3] M. Bohanec, B. Zupan, and V. Rajkovic. Hierarchical multi-attribute decision models and their application in health care. In *Medical Informatics Europe '99*, pages 670–675. IOS Press, 1999.

[4] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10:1895–1924, 1989.

[5] B. S. Everitt. *The Analysis of contingency tables.* Chapman and Hall, London, 1977.

[6] W. H. Harris. Traumatic arthritis of the hip after dislocation and acetabular fractures: Treatment by mold arthroplasty. end result study using a new methods of result evaluation. *J. Bone Joint Surg.*, 51-A:737–55, 1969.

[7] S. E. James and S. C. Gallannaug. Bi-articular hemiarthroplasty of the hip: a 7-year follow-up. *Injury*, 22(5):391–3, 1991.

[8] R. C. Johnston, R. H. Fitzgerald, R. Poss, M. E. Muller, and C. B. Sledge. Clinical and radiographic evaluation of total hip replacement. a standard system of terminology for reporting results. *J Bone Joint Surg [Am]*, 72(2):161–8, 1990.

[9] I. Kononenko, I. Bratko, and M. Kukar. Application of machine learning to medical diagnosis. In *Machine Learning and Data Mining: Methods and Applications*, pages 389–408. John Wiley & Sons, Chichester, 1998.

[10] K. J. Koval, M. L. Skovron, G. B. Aharonoff, and J. D. Zukerman. Predictions of functional recovery after hip fracture in the eldery. *Clin Orthop*, (348):22–28, 1998.

[11] K. J. Koval and J. D. Zukerman. Hip fractures are an increasingly important public health problem (editorial). *Clin Orthop*, (348):2, 1998.

[12] N. Lavrač, S. Džeroski, V. Pirnat, and V. Križman. The utility of background knowledge in learning medical diagnostic rules. *Applied Artificial Intelligence*, 7:273–293, 1993.

[13] J. Lubsen, J. Pool, and E. van der Does. A practical device for the application of a diagnostic or prognostic function. *Methods of Information in Medicine*, 17:127–129, 1978.

[14] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, editors. *Machine learning, neural and statistical classification.* Ellis Horwood, 1994.

[15] R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[16] T. L. Saaty. *Multicriteria decision making: The analytic hierarchy process.* RWS Publications, 1993.

[17] B. Zupan, M. Bohanec, J. Demšar, and I. Bratko. Learning by discovering concept hierarchies. *Artificial Intelligence*, 109(1–2):211–242, 1999.

[18] B. Zupan, J. Demšar, M. W. Kattan, J. R. Beck, and I. Bratko. Machine learning for survival analysis: A case study on recurrence of prostate cancer. In W. Horn et al., editor, *AIMDM-99*, pages 346–355. Springer–Verlag, June 1999.

[19] B. Zupan and S. Džeroski. Acquiring and validating background knowledge for machine learning using function decomposition. *Artificial Intelligence in Medicine*, 14(1–2):101–17, 1998.