Attribute Interactions in Medical Data Analysis

Aleks Jakulin[1], Ivan Bratko[1,2], Dragica Smrke[3], Janez Demšar[1], and Blaž Zupan[1,2,4]

[1] Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, Ljubljana, Slovenia
[2] J. Stefan Institute, Jamova 39, Ljubljana, Slovenia
[3] Dept. of Traumatology, University Clinical Center, Ljubljana, Slovenia
[4] Dept. of Human and Mol. Genetics, Baylor College of Medicine, Houston, USA

Abstract. There is much empirical evidence about the success of naive Bayesian classification (NBC) in medical applications of attribute-based machine learning. NBC assumes conditional independence between attributes. In classification, such classifiers sum up the pieces of class-related evidence from individual attributes, independently of other attributes. The performance, however, deteriorates significantly when the "interactions" between attributes become critical. We propose an approach to handling attribute interactions within the framework of "voting" classifiers, such as NBC. We propose an operational test for detecting interactions in learning data, and a procedure that takes into account the detected interactions in learning. This approach induces a structuring of the domain of attributes, may lead to improved classifier's performance and may provide useful novel information for the domain expert when interpreting the results of learning. We report on its application in data analysis and model construction for the prediction of clinical outcome in hip arthroplasty.

1 Introduction

The most common form of machine learning is attribute-based supervised inductive learning. Given a set of *instances*, each of them described by the values of the *attributes* and the *class*, we learn to predict the class of a new instance. In this paper we consider this learning problem when both the attributes and class are *nominal*. That is, the domains of the attributes and the class are discrete and unordered. Since the class is nominal in our case, the prediction task is classification.

Naive Bayes Classification (NBC) is a machine learning method, particularly popular in medical applications. NBC assumes that the attributes are mutually independent. Although in practice this assumption is not quite true, experience shows that the NBC approach in medical applications is effective and gives relatively good classification accuracy in comparison with other, more elaborate learning methods. Similar conclusions hold for logistic regression, another classification learning method that approximates the target concept as if the attributes were independent. The relative strength of these approaches comes precisely from the fact that they assume attribute independence, even when the assumption is not completely true. The independence assumption licences the classifier to collect the evidences about the class from individual attributes separately. So an attribute's contribution of evidence about the class is determined independently from other attributes. This makes the estimates of evidence from given learning data more robust than in cases when attribute dependences are taken into account. This increase in robustness is particularly important when data is scarce, which is a common problem in medical applications. The evidence from individual attributes can be estimated from larger data samples, whereas the handling of attribute dependences leads to fragmentation of available data and consequently to unreliable estimates of evidence. Often in practice these unreliable estimates cause inferior performance of more sophisticated methods. Consequently, more sophisticated methods (which do not assume independences) often perform inferior to simple NBC or logistic regression.

So NBC often works well in practice, in particular in medical applications, as long as the attributes are "sufficiently independent." However, when attribute dependences become critical, ignoring dependences leads to disastrous performance. Methods like NBC that look at just one attribute at a time are in machine learning called "myopic." Such methods compute evidence about the class separately for each attribute (independently from other attributes), and then simply "sum up" all these pieces of evidence. This "voting" does not have to be the actual arithmetic sum (for example, it can be the product, that is the sum of logarithms, as in NBC). But the important point is that the aggregation of pieces of evidence coming from individual attributes does not depend on the relations among the attributes. We will refer to such methods as "voting methods;" they employ "voting classifiers."

A well known example where the myopia of voting methods results in complete failure, is the concept of exclusive OR: $C = XOR(X, Y)$, where C is a Boolean class, and X and Y are Boolean attributes. Myopically looking at attribute X alone provides no evidence about the value of C. The reason is that the relation between X and C critically depends on Y. For $Y = 0, C = X$; for $Y = 1, C \neq X$. Similarly, Y alone fails. However, *X and Y together* perfectly determine C. We say that there is a *positive interaction* between X and Y with respect to C. In the case of positive interaction between X and Y with respect to class C, the evidence from jointly X and Y about C is greater than the sum of the evidence from X alone and evidence from Y alone.

The opposite may also happen, namely that the evidence from X and Y jointly is worth less than the sum of the individual pieces of evidence. In such cases we say that there is a *negative interaction* between X and Y w.r.t. C. A simple example is when attribute Y is (essentially) a duplicate of X. For example, the length of the diagonal of a square duplicates the side of the square. Similar to positive interactions, voting classifiers are confused by negative interactions as well.

In this paper we propose an approach to handling attribute interactions within the framework of voting classifiers, such as Naive Bayes Classifier. We propose an operational test for detecting positive and negative interactions in learning data, and a procedure for "resolving" the detected interactions when learning a voting classifier. The key in resolving interaction is that the interacting pairs of attributes are treated jointly, giving rise to new attributes, which is similar to the idea of structured induction [1–3]. This approach induces an automatic structuring of the domain of attributes. In addition to improved classifier performance, it is hoped that such domain structuring also provides useful novel information for the domain expert when interpreting the results of learning.

We apply our proposed approach to the medical problem of predicting the success of hip arthroplasty in terms of Harris Hip Score (HHS; [4]). We also compare the automatically induced attribute structure based on interaction analysis, with the structure proposed by a medical expert for the same domain [5].

2 Attribute Interactions

Let us first define the concept of interaction among attributes formally. Let there be a learning problem with the class C and attributes $X_1, X_2, \ldots$. Under conditions of noise or incomplete information, the attributes need not determine the class values perfectly. Instead, they provide some "degree of evidence" for or against particular class values. For example, given an attribute-value vector, the degrees of evidence for all possible class values may be a probability distribution over the class values given the attribute values.

Let the *evidence function* $f(C, X_1, X_2, \ldots, X_k)$ define a "chosen" true degree of evidence for class C in the domain. The task of machine learning is to induce an approximation to function f from learning data. In this sense, f is the target concept for learning. In classification, f (or its approximation) would be used as follows: if for given attribute values $x_1, x_2, \ldots, x_k : f(c_1, x_1, x_2, \ldots, x_k) > f(c_2, x_1, \ldots, x_k)$, then the class c_1 is more likely than c_2.

We define the presence, or absence, of interactions among the attributes as follows. If the evidence function can be written as a ("voting") sum:

$$f(C, X_1, X_2, \ldots, X_k) = g\left(\sum_{i=1,2,\ldots,k} g_i(C, X_i)\right) \tag{1}$$

for some functions g, and $g_1, g_2, \ldots, g_k$, then there is no interaction between the attributes. Equation (1) requires that the joint evidence of all the attributes can be essentially reduced to the sum of the pieces of evidence $g_i(C, X_i)$ from individual attributes.

If, on the other hand, no such functions $g, g_1, g_2, \ldots, g_k$ exist for which (1) holds, then there *are* interactions among the attributes. The strength of interactions IS can be defined as $IS := f(C, X_1, X_2, \ldots, X_k) - g\left(\sum_i g_i(C, X_i)\right)$. IS greater than some positive threshold would indicate a positive interaction, and IS less than some negative threshold would indicate a negative interaction. Positive interactions indicate that a holistic view of the attributes unveils new evidence. Negative interactions are caused by multiple attributes providing the same evidence, which should get counted only once.

We will not refine this definition to make it applicable in a practical learning setting. Instead, we propose a heuristic test for detecting positive and negative interactions in the data, in the spirit of the above principled definition of interactions. Interaction gain is based on the well-known idea of information gain. *Information gain* of a single attribute X with the class C [6], also known as mutual information between X and C, is defined as:

$$\mathrm{Gain}_C(X) = I(X;C) = \sum_{x\in\mathcal{D}_X}\sum_{c\in\mathcal{D}_C} P(x,c)\log\frac{P(x,c)}{P(x)P(c)}, \qquad (2)$$

Information gain can be regarded as a measure of the strength of a 2-way interaction between an attribute X and the class C. In this spirit, we can generalize it to 3-way interactions by introducing the *interaction gain* [7]:

$$IG_3(XYC) := I(\overline{XY};C) - I(X;C) - I(Y;C), \qquad (3)$$

We have joined the attributes X and Y into their Cartesian product $\overline{XY}$. Interaction gain can be understood as the difference between the actual decrease in entropy achieved by the joint attribute $\overline{XY}$ and the expected decrease in entropy with the assumption of independence between attributes X and Y. The higher the interaction gain, the more information was gained by joining the attributes in the Cartesian product, in comparison with the information gained from single attributes. When the interaction gain is negative, both X, and Y carry the same evidence, which was consequently subtracted twice.

3 Attribute Interaction Analysis In Hip Arthroplasty Domain

We have studied attribute interactions and the effect they have on performance of the naive Bayesian classifier in the domain of prediction of patient's long term clinical status after hip arthroplasty. The particular problem domain was chosen for two main reasons. First, the construction of a good predictive model for hip endoprosthesis domain may provide the physician with a tool to better plan the treatment after the operation — in this respect, discovery of interesting attribute interactions is beneficial. Second, in our previous study [5] the participating physician defined an attribute taxonomy for this domain in order to construct a required concept hierarchy for the decision support model: this provided grounds for comparison with the a taxonomy discovered by observing attribute interactions from the data.

3.1 The Data

The data we have considered was gathered at Department of Traumatology of University Clinical Center in Ljubljana from January 1988 to December 1996.

For each of the 112 patients, the data records 28 attributes observed at the time of or immediately after the operation. All attributes are nominal and most, but not all, are binary (e.g., presence or absence of a complication). Patient's long-term clinical status was assessed in terms of Harris hip score [4] at least 18 months after the operation. Harris hip score gives an overall assessment of the patient's condition and is evaluated by a physician who considers, for example, patient's ability to walk and climb stairs, patient's overall mobility and activity, presence of pain, etc. The numerical Harris hip score in scale from 0 to 100 was discretized into three classes: *bad* (up to 70, 43 patients), *good* (between 70 and 90, 34 patients) and *excellent* (above 90, 35 patients).

3.2 Interaction Gain Analysis

We first analyzed the hip arthroplasty data to determine the interaction gain (3) between pairs of attributes. Results of these analysis are presented in Fig. 1, which, for the presentation clarity, shows only the most positive ($IG_3 \geq 0.039$) and the most negative interactions ($IG_3 < -0.007$).

The domain expert first examined the graph with positive interactions; they surprised her (she would not immediately think about these if she would be required to name them), but could all justify them well. For instance, with her knowledge or knowledge obtained from the literature, specific (bipolar) type of endoprosthesis and short duration of operation significantly increases the chances of a good outcome. Presence of neurological disease is a high risk factor only in the presence of other complications during operation. It was harder for her to understand the concept of negative interactions, but she could confirm that the attributes related in this graph are indeed, as expected, correlated with one another. In general, she found the graph with positive interactions more revealing and interesting.

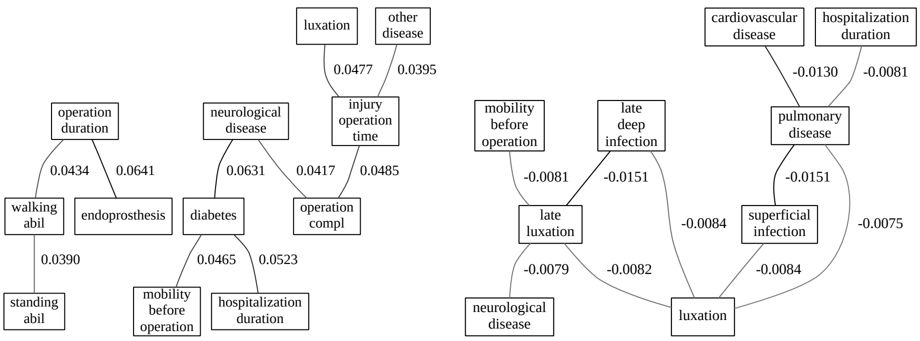

Fig. 1. Graphs displaying the distinctly positive (the two subgraphs on the left), and negative (the graph on the right) interactions. Each edge is labeled with the value of IG_3 for the pair of connected attributes.

3.3 Induction of Attribute Structure

To further observe interactions in our domain, we used the hierarchical clustering method 'agnes' [8]. Pairs of attributes that interact strongly with the class, either positively or negatively, should appear close to one another, while those which do interact should be placed further apart. They do not interact if they are conditionally independent, which also happens when one of the attributes is irrelevant. The dissimilarity function, which we express as a matrix D, was obtained with the following formula:

$$\mathsf{D}(A,B) = \begin{cases} |1/\mathrm{IG}(ABC)| & \text{if } |IG(ABC)| > 0.001, \\ 1000 & \text{otherwise.} \end{cases} \tag{4}$$

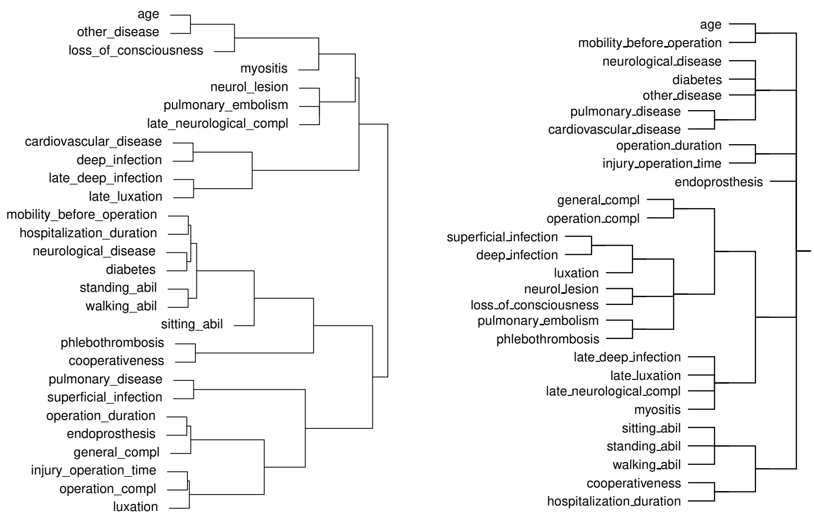

Fig. 2. An attribute interaction dendrogram (left) illustrates which attributes interact, positively or negatively, while the expert-defined concept structure (right) was reproduced from [5].

In Fig. 2, we compared the attribute interaction dendrogram with an expert-defined concept structure (attribute taxonomy) that was used as a skeleton for decision support model in our previous study [5]. While there are some similarities (like close relation between ability to stand and walk), the two hierarchies are mostly dissimilar. The domain expert appears to have defined her structure on the basis of medical (anatomical, physiological) taxonomy; this seems not to correspond to attribute interactions, as we have defined them in this text.

4 Construction of Classification Models

While the naive Bayesian classifiers cannot exploit the information hidden in a positive interaction [9, 10], the attributes in negative interactions tend to confuse their predictions [11]. The effects of negative interactions have not been studied extensively, but provide explanation for benefits of feature selection procedures, which are one way of eliminating this problem.

With *resolving interactions*, we refer to a procedure where the interacting pairs of attributes are treated jointly, giving rise to new attributes which are added to the data set. The best subset of attributes is then found using a feature subset selection technique, and later used for construction of a target prediction model. For feature subset selection, we used the greedy heuristic, driven by the myopic information gain (2): only the n attributes with the highest information gain were selected. For resolution of interactions we also used a greedy heuristic, guided by the interaction gain (3): we introduced the Cartesian product attributes only for the N attribute pairs with the highest interaction gain.

In our experimental evaluation, interaction gain scores were obtained from considering the complete data set, new attributes were generated, and added into the data set. In the second phase, the naive Bayesian classifier, was built using the altered data set and evaluated at different sizes of the selected feature subset. The ordering of the attributes for feature subset selection using information gain, and the modeling using the subset were both performed on the learning data set, but evaluated on the test set. The evaluation was performed using the leave-one-out schema: for the data set containing l instances, we performed l iterations, $j = 1, 2, \ldots, l$, in which all instances except j-th were used for training, and the resulting predictive model was tested on the j-th instance. We report average performance statistics over all l iterations. All the experiments were performed with the Orange toolkit [12].

To measure the performance of classification models we have used two error measures. *Error rate* is the proportion of test cases where the classifier predicted the wrong class, i.e., the class for which the classifier predicted the highest probability was not the true class of the test case. The second error measure, *Brier score*, is usually used to assess the quality of weather forecasting models [13, 14], and recently gaining attention in medicine [15]. It is better suited for evaluating probabilistic classifiers because it measures the deviations from the actual to the predicted outcome probabilities. As such, it is more sensitive than the error rate. A learning method should attempt to minimize both the error rate and the Brier score.

We have assessed how the inclusion of different number of newly constructed and original attributes affects the prediction performance. Figure 3 illustrates the search space for our particular domain, where the number n of attributes selected is plotted on the horizontal and the number N of interactions resolved on the vertical axis. The best choice of n and N can be determined with a wrapper mechanism for model selection. We can observe several phenomena: increasing the number of attributes in the feature subset does not increase the error rate as much as it hurts the precision of probability estimates, as measured by the

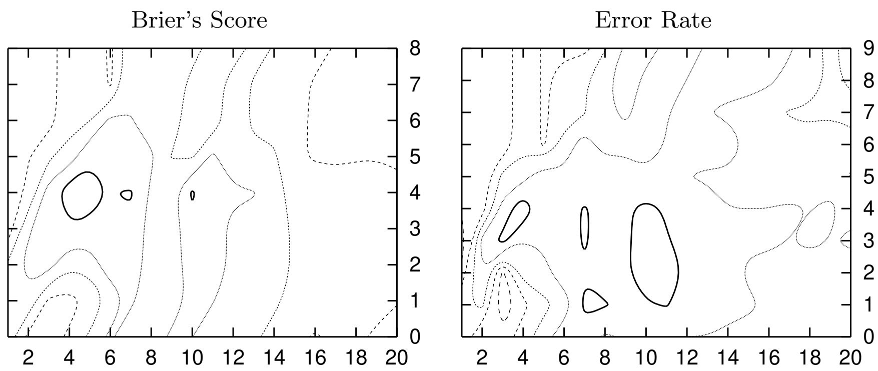

Fig. 3. Dependence of the Brier score and error rate on the feature subset size, n (horizontal axis) and on the number of interactions resolved, N (vertical axis). Emphasized are the areas of the best predictive accuracy, where Brier score is less than 0.2 and the error rate less than 0.45.

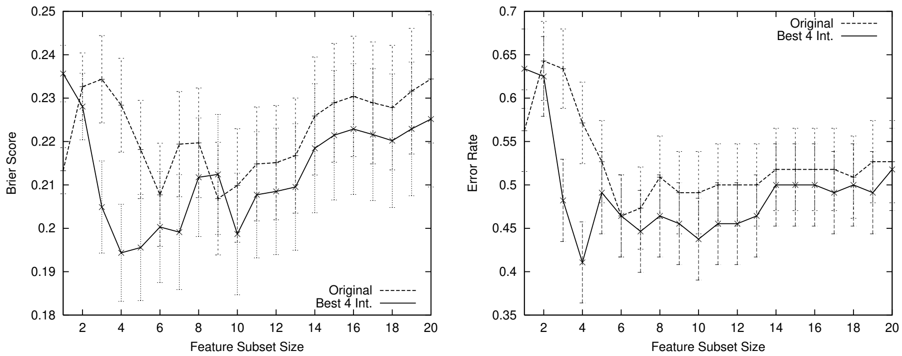

Fig. 4. Average Brier score and error rate as computed by leave-one-out and its dependence on the number of attributes used in the model for $N = 4$ (solid line) and $N = 0$ (dashed). For all measurements, the standard error is shown.

Brier score. Furthermore, there are diminishing returns to resolving an increasing number of interactions, as illustrated in the contour diagrams in Fig. 3. Unnecessary interactions merely burden the feature subset selection mechanisms with additional negative interactions, because attributes that arise out of interaction resolution contain the information already provided by the original attributes. Figure 4 presents the results in terms of Brier score and error rate with four resolved interactions.

There are several islands of improved predictive accuracy, but the best appears to be the area with approximately 4 resolved interactions and 4 selected attributes. Classification accuracy reaches its peak of 59% at the same number of attributes used. This accuracy improves upon the accuracy of 56% obtained in our previous study, where manually crafted features as proposed by domain experts were used in the naive Bayesian classifier [5]. Both are a substantial improvement over models constructed from the original set of features, where the accuracy of NBC with the original 28 attributes is 45%, and does not rise beyond 54% even with the use of feature subset selection. The results in Table 1 show that three of the four constructed attributes were chosen in building of the model. The table provides the set of important interactions in the data, where an important increase in predictive accuracy can be seen as an assessment of the interaction importance itself, given the data.

Table 1. Average information gain and feature selection rating for attributes for the case $N = 4, n = 4$. The resolved interactions are emphasized.

Information Gain	Attribute
0.118	**luxation + injury_operation_time**
0.116	**diabetes + neurological_disease**
0.109	**hospitalization_duration + diabetes**
0.094	pulmonary_disease

5 Summary and Conclusions

We have defined interactions as deviations from the independence assumption between attributes. Positive interactions imply conditional dependence of attributes given the class; new evidence is unveiled if the positively interacting attributes are treated jointly. Negative interactions imply mutual dependence of attributes and duplication of evidence; we must be careful not to account for the same evidence more than once. We have introduced interaction gain as a heuristic estimate of the interaction magnitude and type for 3-way interactions between a pair attributes and the class.

We have proposed a method for analysis and management of attribute interactions in prognostic modeling. In an experimental evaluation on hip arthroplasty domain, we have obtained a number of promising and unexpected results. Promising were those based on performance evaluation: resolution of positive interactions yielded attributes that could improve the performance of predictive model built by the naive Bayesian classification method. Promising but also unexpected were the interactions themselves: we have observed that pairs of interacting attributes proposed using our algorithm and induced from the data were quite different from those obtained from expert-designed attribute taxonomy. Although the new attributes proposed by experts can constitute a very valuable part of a background knowledge, and may significantly improve the performance of predictive models (see [5]), other important attribute combinations may be overlooked. The algorithms described in this paper may help the domain experts to reveal them and, if found meaningful, include them in their knowledge base.

As any new method, the one proposed in this paper requires further evaluation on different problems and data sets. Beyond the sole test case reported here, we did perform various other types of preliminary analysis [7]. They all pointed to potential usefulness of the method, most particularly for *feature subset selection* that is particularly tailored for methods such as the naive Bayes and logistic regression, *constructive induction* where interaction gain can guide the selection of appropriate attribute sets for which new features should be constructed, and for *data analysis*, where appropriate visualization techniques, such as those presented in this paper, may help domain experts and data miners to gain further insight into interplay of attributes, their role and importance in a predictive model.

References

1. Shapiro, A.D.: Structured induction in expert systems. Turing Institute Press in association with Addison-Wesley Publishing Company (1987)
2. Michie, D.: Problem decomposition and the learning of skills. In Lavrač, N., Wrobel, S., eds.: Machine Learning: ECML-95. Notes in Artificial Intelligence 912. Springer-Verlag (1995) 17–31
3. Zupan, B., Bohanec, M., Demšar, J., Bratko, I.: Learning by discovering concept hierarchies. Artificial Intelligence **109** (1999) 211–42
4. Harris, W.H.: 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 **51-A** (1969) 737–55
5. Zupan, B., Demšar, J., Smrke, D., Božikov, K., Stankovski, V., Bratko, I., Beck, J.R.: Predicting patient's long term clinical status after hip arthroplasty using hierarchical decision modeling and data mining. Methods of Information in Medicine **40** (2001) 25–31
6. Hunt, E.B., Martin, J., Stone, P.: Experiments in Induction. Academic Press, New York (1966)
7. Jakulin, A.: Attribute interactions in machine learning. Master's thesis, University of Ljubljana, Faculty of Computer and Information Science (2003)
8. Struyf, A., Hubert, M., Rousseeuw, P.J.: Integrating robust clustering techniques in S-PLUS. Computational Statistics and Data Analysis **26** (1997) 17–37
9. Kononenko, I.: Semi-naive Bayesian classifier. In Kodratoff, Y., ed.: European Working Session on Learning - EWSL91. Volume 482 of LNAI., Springer Verlag (1991)
10. Domingos, P., Pazzani, M.: On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning **29** (1997) 103–130
11. Rish, I., Hellerstein, J., Jayram, T.: An analysis of data characteristics that affect naive Bayes performance. Technical Report RC21993, IBM (2001)
12. Demšar, J., Zupan, B.: Orange: a data mining framework. http://magix.fri.uni-lj.si/orange (2002)
13. Brier, G.W.: Verification of forecasts expressed in terms of probability. Weather Rev **78** (1950) 1–3
14. Atger, F.: The skill of ensemble prediction systems. Monthly Weather Review **127** (1999) 1–13
15. Margolis, D.J., Halpern, A.C., Rebbeck, T., et al.: Validation of a melanoma prognostic model. Arch Dermatol. **134** (1998) 1597–1601