Collaborative Network of Predictive Markers Complicates Formation of Prognostic Groups in Patients with Advanced Lung Cancer

BERNICE SCHWAB1, ANDREAS RUDOLPH2, EDGAR JOST3 and UWE KLINGE1

1Department of General, Visceral and Transplant surgery, University Hospital of the RWTH Aachen, Aachen, Germany; 2Universität der Bundeswehr München, Institut für Mathematik und Informatik FB BW Werner-Heisenberg-Weg, Neubiberg, Germany; 3Department for Oncology, Hematology and Stem Cell Transplantation, Medical Faculty, RWTH Aachen University, Aachen, Germany

Abstract. Background: Evaluation of cancer therapies is mainly based on prolonging remission status and effect of survival. Various serological, clinical or histological markers are used to estimate the patient’s prognosis, and to tailor specific therapies for patients with poor prognosis. However, it is still a challenge to combine all this information into a comprehensive risk prediction. Patients and Methods: In 58 patients with advanced non small cell lung cancer we recorded 38 parameters (15 from clinic, 10 from histology, 13 from serology) to analyze their impact on survival. We both used univariate as well as multivariate approaches and decision tree analysis. Results: Univariate analysis showed that ECOG status, stage, and the presence of cerebral or bone metastasis had a significant impact on survival, as well as the serum markers CA15-3, TPA, Cyfra. In a multivariate approach only ECOG and stage had a significant impact on survival. Considering correlation coefficients of >0.3 as an indicator of a functional relationship, we found several relations among the clinical (9), histological (8) or the serological parameters (13). Survival was related to 9 parameters by significant direct and cross-relation coefficients. The use of already few variables with its different possible options led to many different patterns in the cohort, almost all being specific for individual patients, and thereby underlining their heterogeneity. Decision tree analysis revealed that by including either stage and kind of therapy or stage and expression of YB-1 allows to identify sub-groups with distinct prognosis. Conclusion: Clinical, serological and histological markers, all provide prognostic information. Because they are all linked in a collaborative network, the formation of homogenous prognostic groups by use of single markers is limited. Alternative statistical approaches with focus on decision trees may allow use of various information to assess individual patients into distinct risk groups.

Cancer regularly represents a mixture of different sub-groups rather than a unified entity with a defined prognosis; this fact has been reinforced by our increasing skills for genetic and proteomic profiling. Theodor Billroth, one of the most prominent surgeons of the 19th century suspected that in a cohort of 1,000 patients only 25-100 patients may show a similar course, and already in 1886 he demanded for a prognostic coefficient to identify such a group of patients with comparable outcome (1).

For identification of optimum therapy it is necessary to identify variable(s), which provide prediction of outcome prior to starting treatment. Whereas experience of the past decades had clearly confirmed the importance of any clinical staging for selecting an appropriate therapeutic strategy, it has become apparent that molecular variability of the malignancy has also an impact on prognosis. In recent years several different genes, proteins and clinical information have been evaluated for their predictive value. Regarding the huge heterogeneity of patients with a malignancy, differing in stage of the tumor, cellular and molecular patterns of the tumors, co-morbidities or personal history of past treatments there seems to be no sole indicator, which is able to sufficiently consider all these differences. Correspondingly, attempts for grouping with regard to progression-free survival or to response to therapies are going to be based on combinations of markers and/or clinical information forming comprehensive risk scores. However, the
crucial question becomes apparent, which variables should be included into these scores?

Unfortunately a simple addition of more indicators will not solve the problem, as it is a principal dilemma, which already has been named as “principle of uncertainty” by Grenander in 1951 (10) or as “bias/variance dilemma” by Gemann in 1992 (8). It is based on the fact that any increasing number of variables admittedly helps to describe the constellation of any possible condition with less bias; however, it will add the variability of every single variable and thereby decrease the predictive power markedly. In this regard, variables often do not help to extract patients with similar prognosis but in contrast separate any cohort into exponentially rising numbers of sub-groups. As this diversification eventually may end up in individual cases it even questions any formulation of standard therapies in guidelines, as it rather needs tailored approach towards treatment procedures that are individualized for patients with (on a molecular basis) comparable tumors.

The purpose of the present study, including 58 patients with advanced lung cancer, was to demonstrate the principal challenge for evaluation of various informative markers extracted from genes, proteins, and clinics for its prognostic impact. All markers were taken at the time of the diagnosis and were tested with regard to their impact on prognosis, either as sole markers or integrated in a combination. Even at this small cohort the importance of clinical staging and the decision for the primary therapy could be confirmed. However, though some serological tumor markers indicate a poor prognosis, the addition of tumor cell proteins and various serological markers clearly demonstrate the heterogeneity and individuality of every patient, and thus each of these markers provides only a limited contribution for the prediction of outcome.

Materials and Methods

Patients. We retrospectively investigated the course of 58 patients, treated for advanced NSCLC between 1997 and 2005. All together 38 parameters were measured at the time point of first admittance or first excision of tumor tissue. Patients were treated with chemotherapy for either advanced cancer or relapse of an initially limited cancer, and patients all had biopsies available at the time point of first diagnosis of lung cancer. In 2011 and 2012 we contacted the patient’s physician for information about survival.

Clinical data included tumor grading according to TNM, performance status ECOG (18), stage of Mountain (16), presence of cerebral or bone metastasis, histology, smoking, gender, and age (Table I). Immunohistochemical assessment of the tumor tissue included staining for YB-1, PAI, COX-2, EGGF-Receptor, Notch-3, HIF 1a, Jagged, and p53. Initial values of the following tumor markers were evaluated: CA15-3 (<32 kU/l), CA125 (<36 kU/l), NSE (<14 μg/l), TPA (<92 U/l), CYFRA 21-1 (<2.4 μg/l), CA 72-4 (<7.0 kU/l), CA 19-9 (<38 kU/l), CEA (<5 μg/l), SCC (<1.6 μg/l), PSA (<4 μg/l), β2-Mikroglobuline (<1.8 mg/l), alpha-fetoprotein (<10 μg/l), and CRP (0-46 ng/l). Values of serum markers have been recoded as ratio of value/norm value, with ratio of 1=normal, and every ratio >1 is considered as pathological.

Immunohistochemistry. Tumor samples were evaluated for histology and protein expression from three different experts, blinded and independently. Evaluation was performed separately for cytoplasmatic and nuclear expression. The expression of immunohistochemical parameters at each six different sections were classified by a modified immunoreactive score (IRS) (19). For characterization of the tumor-host interaction the following antibodies were used: CD68 mouse monoclonal antibody (Dako), Notch3 polyclonal anti-goat antibody (Santa Cruz), Cox2 polyclonal rabbit antibody (DICS Innovative Diagnostic Systems). As secondary antibody we used biotinylized goat anti-rabbit for Cox2 and MMP2, goat anti-mouse for CD68, and rabbit anti-goat for Notch3 (all Dako).

Statistical analysis. Statistical evaluation was performed with IBM SPSS statistics 20, using calculation of mean values for cardinal scaled parameters, comparison of mean numerical values by t-test, comparison of distributions in case of nominal scaled values by chi-squared, comparison of ordinal scaled values by univariate and multivariate ANOVA with PostHoc Bonferroni adjustment, calculation of Spearman’s correlation coefficient, calculation of ROC-curves, Kaplan Meier estimates with Log rank comparison. p-Values <0.05 were considered significant. The calculation of the regression tree was performed with the program package R, the data import into R was prepared with LibreOffice Calc (building a so called.csv file). For the mathematical details of a regression tree see Breiman (1984) (2), Tutz (2000) (20) and Zhang u. Singer (2010) (23).

Results

The cohort of the present study consisted of 49 men and 9 women at a mean age of 64 and 61 years respectively. Squamous cell carcinoma was only seen in men (p<0.05). Stages did not differ between gender. p=0.097. Only two
patients were non-smokers. In this cohort of mainly advanced tumor stages 51 patients died, 37 of them within the first year. Only 6 patients were still alive (Figure 1A).

**Modeling survival.** In an univariate linear regression model none of the variables showed a significant ($p<0.05$) impact on survival, neither the clinical nor the tissue or the serum variables. However some of the variables showed a significant correlation of Spearman’s correlation coefficient with survival and a significant effect in the log-rank Mantel-Cox test ($p<0.01$; Figure 1B-E) for ECOG (Spearman’s $r=-0.465$), stage (Spearman’s $r=-0.544$), cerebral (Spearman’s $r=+0.340$) or bone metastasis (Spearman’s $r=+0.368$). Primary therapy reached a correlation coefficient of 0.226 ($p=0.097$).

Out of the serum markers only CA 15-3 ($p=0.002$), TPA ($p=0.001$), Cyfra ($p=0.000$) and CA 72-4 ($p=0.028$) showed a significant correlation with survival, however not a significance in the log-rank test after having recorded the values in a binary digit as normal or pathological. None of the values for expression of the tissue proteins was related significantly with survival. In a multivariate regression model for prediction of outcome with ECOG, stage, cerebral or bone metastasis, CA 15-3, TPA, Cyfra, and CA72-4 again it was only ECOG that contributed significantly ($p=0.04$) to the prediction of survival time.

**Modeling risk for death.** Focusing on death as end-point there was a significant relationship to stage with a significant correlation of $r=0.463$. Whereas 50% of the patients without positive lymph nodes survived, this was reduced to 20% in cases of positive lymph nodes, and dropped to zero in cases of metastasis. Furthermore, ECOG showed a significant correlation ($r=-0.360$, $p=0.008$) with death. Only patients with ECOG 1 or 2 were still alive. Obviously both stage and ECOG were of major relevance for prognosis. Calculation of receiver operator characteristic curves (ROC-curves) for prediction of death furthermore revealed an area under the curve (AUC), which served as indicator of the suitability of a criteria to predict an effect, of more than 0.8 for ECOG (0.878) and stage (0.911; Figure 1F).

**Collaborative network of interferences between variables.** Spearman correlation coefficient was considered as reflection of functional linkage between variables, and interactions were considered as possibly relevant if $r\geq 0.3$.

Among the clinical variables (Figure 2A) we thereby observed that a) an increase of age was related to more $\beta 2$-microglobulin ($r=0.33$), to a higher grading of the tumor ($r=0.32$), but to a lower level of CRP ($r=0.42$). b) Females had smaller tumors ($r=-0.31$) with less $\beta 2$-microglobulin ($r=-0.32$), showed a predominance of adenomateous cancer ($r=-0.33$), and had lower level of CYFRA ($r=-0.40$). c) A lower general condition with higher ECOG was related to higher levels of TPA ($r=0.52$), Cyfra ($r=0.45$) or CRP ($r=0.41$), with higher stages ($r=0.38$), presence of metastasis ($r=0.36$), and shorter survival time ($r=-0.47$). d) Non-smoking was related to longer survival time ($r=0.33$), with less differentiated histology ($r=0.33$) and lower stages ($r=-0.34$). e) A longer survival time was related to absence of metastasis ($r=-0.46$), non-smoking ($r=0.33$), lower levels of Ca72-4 ($r=-0.32$), CA 15-3 ($r=-0.42$), Cyfra ($r=-0.49$) or TPA ($r=-0.50$), with lower ECOG ($r=-0.47$), and lower stages ($r=0.55$). f) Higher stages were of course related to higher levels of T ($r=0.35$), N ($r=0.46$), M ($r=0.88$), but also with higher ECOG ($r=0.38$), smoking ($r=-0.37$) and reduced survival time ($r=-0.55$). g) Squamous cancer was related to expression of PAI in the cytoplasm ($r=0.42$), and with higher levels of CYFRA ($r=0.38$) or SCC ($r=0.37$); they were not found in females ($r=-0.33$). h) A non-differentiated histology was related to higher levels of CYFRA ($r=0.35$), non-smoking ($r=0.33$), lower age ($r=-0.32$), and lower levels of SCC ($r=0.33$).

Among the group of tissue variables (Figure 2B) we found that a) PAI was expressed more often in squamous cancer ($r=0.42$), b) EGF-receptor particularly in patients with cerebral metastasis ($r=-0.31$) and with high levels of CA 72-4 ($r=0.34$). c) Jagged showed a positive relation to higher level of CRP ($r=0.36$). d) Nuclear YB-1 was found in small tumors ($r=-0.34$). e) P53 at advanced tumor stages that could not be treated by surgery ($r=-0.33$). f) Nuclear Notch was expressed in patients with high levels of SCC ($r=0.33$) or CEA ($r=0.32$). Interactions within the group of tissue markers revealed only sole relationships without forming any network-like configuration.

The various serum markers formed a simple but cross-linked network (Figure 2C), in which a) CA 15-3 correlated with CA 125 ($r=0.48$), TPA ($r=0.39$) and Cyfra ($r=0.38$), but was also related to strong expression of COX in the tissue ($r=0.32$), was higher in the presence of cerebral metastasis ($r=-0.31$) and was related to shorter survival time ($r=-0.42$). b) CA 125 was related to CA 15-3 ($r=0.48$), CRP ($r=0.37$). c) CA 72-4 ($r=0.34$), whereas c) NSE was related to bone metastasis ($r=-0.49$) d) TPA was positively related to CYFRA ($r=0.81$), CA 15-3 ($r=0.39$) and NSE ($r=0.36$), to higher stages ($r=0.53$), higher ECOG ($r=0.52$), larger tumor size ($r=0.33$), presence of bone metastasis ($r=-0.53$) and shorter survival time ($r=0.50$). e) Cyfra showed positive correlations with TPA ($r=0.81$), CA 15-3 ($r=0.38$), CA 19-9 ($r=0.32$), and NSE ($r=0.31$), was higher in non-differentiated tumors ($r=0.35$), bigger tumors ($r=0.44$) and reduced survival time ($r=-0.49$). f) CA 72-4 was positively-related to expression of EGF-receptor ($r=0.34$), tumor size ($r=0.32$) and to the presence of cerebral metastasis ($r=-0.30$); it was negatively-related to survival time ($r=-0.32$). g) $\beta 2$-microglobulin was related to the presence of metastasis...
Figure 1. A: Survival curve for 58 patients with advanced NSCLC. B-E: Kaplan Meier curve for survival in relation to (B) ECOG 0 to 3, (C) stage 1 to 4, (D) cerebral filiae, (E) ossary filiae; always p<0.01 in Log-rank Mantel-Cox test. F: Receiver operating characteristic curve for ECOG (- -) and stage (…..) to predict death at 58 patients with NSCLC. AUC for ECOG and stage was 0.878 and 0.911, respectively.
(r=0.34), elevated at higher ages (r=0.33), and in men (r=–0.32). h) CEA was related to the expression of Notch in the cytoplasm (r=0.32). i) CRP was related to poor general condition and higher ECOG (r=0.41), to higher levels of CA 125 (r=0.37) but lower levels of PSA (r=–0.32), was elevated at lower ages (r=–0.42) or in tumors expressing Jagged in the cytoplasm (r=0.36). Whereas many observed correlations in this small cohort reflect common clinical knowledge, survival was significantly influenced by only a few, as mentioned above. However, all these impact factors showed significant positive correlations in between each other, which indicated the close functional cross-linkages (Figure 2D).
Thus, putting all available markers together in a unified comprehensive score would lead to a marked overrating as several markers may reflect similar risks.

Individual marker constellation. Despite a significant impact on the outcome of the patients could be proven for several variables, this was limited to statistics for the entire cohort, and did not necessarily reflect every individual pattern. When using the four clinical variables: gender (binary coding with 1=male, 2=female), stage (coding with 4 levels), histology (binary coding with 1=adeno and 2=squamous), and ECOG (coding with 5 levels 1-5), all together 80 different patterns could be expected, of which in fact 22 were realized in the 53 patients. Sorting from frequent to rarest ended up in a potential course \( y = 21.406x^{-0.8225}, R^2=0.9102 \). The most frequent constellation with the formula “1412” was seen in 9 patients (17%). In case an equal and independent chance for every variable was supposed, every possible pattern should be expected in 1.6% of the cases, which obviously was not the case, and confirmed the cross-linkages among variables (Figure 3A).

In case of the 13 different serum markers recoded into dichotomic values (1=non-pathological, 2=pathological), there should have been \( 2^{13}=8192 \) different possibilities. In our cohort of 58 patients we got an almost individualized distribution with unique pattern for 54 patients (1.7%), and only one constellation that was realized in 4 patients.

Figure 3. Frequency of the different patterns of (A) gender, stage, histology and ECOG or of (B) 13 serum markers recoded as binary digit, sorting in decreasing order, cohort of 53 patients with NCSLC.
Correspondingly, the use of all 13 serum markers led to individual constellations and did not permit formation of groups (Figure 3B).

**Regression tree analysis**

*Stepwise forward.* With the given clinical variables and survival time as dependent variable we got the following rules (Figure 4A): 1. If stage=4 or missing, the mean survival time was equal 8.3. 2. If stage=1, 2 or 3 and if primary therapy was radiochemotherapy, the mean survival time was 17.8 months. 3. If stage=1, 2 or 3 and if primary therapy was only chemotherapy or included surgery (coded with 1, 4 or 5), the mean survival time was 63.4.

Correspondingly, the following variables built up the regression tree: “stage” and “primary therapy”.

As one can see from Figure 4B if stage=4, this group contained 34 patients, whereas if stage=1, 2 or 3, the corresponding group contained 24 patients.

If one splits the second group by primary therapy=3, then the resulting group contained 11 patients, whereas if primary therapy=1, 4, or 5, then the resulting group contained 13 patients (11+13=24).

With all available variables as independent variables and survival time as dependent variable we have the following rules: 1. If YB1 expression with a maximum of 20 in the cytoplasm (YB1_cyto)=10, 11, 12, 13, 14, 15, 17, 19, 3, 7, or 8 and if stage=3 or 4 then survival time was in the mean 8.7 months. 2. But with a similar YB1 expression in the cytoplasm and a stage =1 or 2 the mean survival time was 42.3 months. 3. If YB1 expression in the cytoplasm was found to be=16, 20, 4, 6 or 9 then survival time was 59.7 months.

Consequently the regression tree was built-up by the variables “stage” and distinct values of “YB1 expression” in the cytoplasm.

*Stepwise backward.* Taking all available variables as independent variables and survival time as an dependent variable and leaving out variables from the group of the immunohistochemical variables one after the other led again to a model with “stage” and “primary therapy” as most important independent variables.

**Discussion**

In the present study, on 58 patients with advanced NSCLC the clinical information given by ECOG, stage or the knowledge of metastasis is of major importance for the outcome of the patients. This was confirmed by univariate, multivariate and correlation analysis. In contrast, measurement of protein expressions in tumor tissue could not demonstrate any significant correlation to outcome, whereas elevation of three of the 13 serum tumor markers indicated a poor survival (Ca 15-3, TPA, Cyfra).

Although many markers did not show a significant impact on outcome, correlation analysis revealed numerous intercorrelations to other markers, which all seemed to form a collaborative network. Despite the fact that we are far from having understood the details of all the complex interactions, it became clear that a sole and unique marker is unlikely to be sufficient to reflect all the prognostic impacts. However, any combination of markers has to consider that the functional inter-relationship among them will lead to an overestimate, when closely related markers were included in one comprehensive score.

**Impact factors used for prediction of outcome.** In our study clinical information as ECOG and stage were found to be of
major relevance for the outcome of the patients. This is in accordance with experience of Kuo et al., who proposed a simplified co-morbidity score (SCS) (13). Patients with an SCS >9 had shorter overall survival than those with SCS ≤9 both in limited-stage (372 days versus 581 days, p=0.01) and extensive-stage disease (215 days versus 324 days, p=0.001). Multivariate analysis indicated that SCS >9 was associated with a worse prognosis in patients with limited-stage disease (HR 2.17, 95% CI 1.12-4.21) and extensive-stage disease. However, the basis for selection or weighting of the variables has not been outlined.

In contrast to the importance of clinical data we did not find any new significant correlation between expression of marker proteins and survival. However, we observed an enormous variation for every tissue marker, covering the entire range of 0 to 20 with almost every marker. It might be this heterogeneity, which hinders any detection of close relationship. A similar variation has been described by Hu et al., who studied the expression patterns of USP22 and potential targets BMI-1, PTEN, p-AKT in non-small-cell lung cancer. Combining the 4 markers resulted in an independent prognostic indicator of overall survival (p<0.001; HR 5.974; 95%CI 3.307-10.791) (11). However, the results reflect that with each marker there were always some negative, and that there was no uniform pattern. It is questionable whether a significant impact of such a prognostic indicator may be preserved if further markers are added. Liu considered not only the intensity of the expression of FOXP3 and of CD8 but identified the ratio between them as prognostic relevant (14). They found higher FOXP3(+) /CD8(+) ratio particularly in tumor sites of patients with poor response to platinum-based chemotherapy. These results indicate that not only the expression of a specific marker but that additionally the local defense mechanism has to be considered as balance of pro and con effects, some improving outcome whereas others not. This need for balancing the impact increases the complexity even more, regardless, whether a risk score uses information from clinical marker, from serum analysis or tissue investigations.

The combination of clinical and genetic information can add predictive information compared to the impact of the sole marker. De Fraipont investigated the 3p allelic balance by defining an apoptosis methylation prognostic signature, and could identify three sub-groups with strikingly different prognosis using a translational approach including the information of DAPK1 methylation, tumor stage, and RASSF1A methylation (5).

However, it may be an illusion that the challenge of increased variance by expanding the number of variables can be met by genetic profiling. Chen et al. performed an extensive genetic analysis to define a malignancy-risk gene signature using a large NSCLC microarray dataset and two independent NSCLC microarray data sets (3). They defined a risk score from 67 statistically significant malignancy-risk genes, and found that those with a low malignancy-risk score had increased overall survival (OS) compared to a high malignancy-risk. Interestingly, the genetic risk score was independent from TNM stage, indicating that any stage includes patients with poor and good prognosis. Overall, TNM as well as the malignancy risk score had both a significant impact on survival. But even with the help of these huge data sets it was not possible to give an individual risk profile without integration of clinical information. Dai et al. analyzed 178 SNPs from 52 immune genes (4). Combination of four SNPs with stage and surgery resulted in the best prediction with the lowest number of variables. In another study Wan identified a total of 21 gene signatures supporting prediction of patients’ outcome. In multivariate analyses with age, gender, race, smoking history, cancer stage, and tumor differentiation, a 10-gene signature had a hazard ratio of 3.23 (95%C I 1.48, 7.06), which was a more significant prognostic factor than other clinical factors, except cancer stage (stage II and stage III) (21). Accuracy in a training set, and two evaluations sets were 64%, 57% and 66%, respectively. However, this study did not address the problem of overfit when combining genes, which were closely co-expressed. Furthermore the authors used a 10-gene score though they had identified 21 genes as relevant.

Variability of many parameters challenges comparability of patients. As delineated above numerous data have been collected and evaluated with regard to their importance in characterizing the tumor, however, a typical challenge remains a correct analysis of big numbers of variables. Gigarenzer clearly outlined that an increasing number of variables not only improves the possibility to model an outcome, but enhances the variance and the number of possible patterns exponentially (9). Wu et al. analyzed the impact of PXN and of miR-218 (22). After recoding as binary digit they had to consider 4 possible combinations of these 2 markers and found worst OS and RFS only for the condition with low miR-218 combined with positive PXN. Already 10 markers with a binary coding will result in 2^10 different patterns, and with 20 markers we have to consider more than a million different patterns. Consideration of all genes and proteins of a patient will end up in a figure far larger than the number of all atoms in the universe. Obviously, the mere accumulation of a complete set of information of all genes and proteins will not help us.

Definition of 38 markers for the 58 patients of this study ends-up in individual patterns with unique conditions for each patient. A cohort size of 50 or even 5,800 really would not have changed anything of relevance facing 2^38=more than 274 billion different options for 38 variables. Correspondingly, high number of variables will separate even enormous cohorts into cases of individuals, e.g. as done by
Algorithms for an individual risk score. As we have to face several survival supporting or life-shortening effects any individual risk score has to reflect the personal balance, and thus gives an advice for the best treatment. It may be necessary to focus on three distinct groups: Those with excellent outcome, those with poor outcome and those which as intermediate can not be located in one or the other. In this regard several attempts have been discussed, like a prognostic index for second-line chemotherapy by Di Maio, indicating a good discrimination to the proposed three risk categories (7).

The baseline model presented by Dehing-Oberhije analyzed clinical data together with several blood biomarkers, and yielded an AUC of 0.81 (6). Lopez-Encuentra et al. similarly studied on early-stage NSCLC, the benefit of integrating pathological TNM, clinical and molecular factors into a composite prognostic model (15). Combining all variables the AUC increased to 0.74. “The model of the integrated group classified patients with significantly higher accuracy compared to the TNM 2010 staging.”

Any individual score has to consider all factors with relevant impact on outcome, but to avoid any overestimate should not count them in double or even more, when the variables may contain information of subsequent levels and thus are linked between each other. An appropriate weighting is even more difficult for variables, which showed a non-linear relationship with outcome (17). Non-linear relationships either u-shaped or even more complex makes it very difficult to integrate this information into disease models as it generally requires a switch from binary to a variable’s coding of at least 3 levels. Correspondingly the number of different constellations increases by $3^n$ variables, which means already more than 59,000 possible constellations in case of only 10 variables.

One approach to estimate the individual balance might be the score described by Kössler et al. where the formation is based on correlation coefficients for predicting outcome in patients with NSCLC. By combining protein expression analysis of CD68 and GAS6 with T, N and M, using either Cox regression or ISIR, they could improve prediction of outcome (12).

Regression tree to tailor therapies. In the current study we performed regression analysis to develop an algorithm with prognostic impact, which can be used as decision criteria for individual patients. Essentially such a tree is built-up by partitioning the sample space (i.e. splitting the sample space into disjoint sets) such that on every subset the dependent variable is constant. This is done with the help of some goodness of fit criterion (typically the variance as forecast criterion for the dependent variable), which lead to a description of such a partition by rules, i.e. in this study we used the parameter of survival time.

This methodology has some advantages compared with other statistical methodologies: • There is no assumption with respect to a functional structure as for example in case of an analysis of variance. • There is no assumption like a normal distribution. • The independent variables can be of quantitative nature as well as of qualitative nature. • The scale of the variables does not matter. • The curse of dimensionality does not matter.

Due to the generating process, partitioning leads to a tree structure, more precisely a binary tree. In case of a quantitative variable a decision is taken for the left or the right node, respectively, if the considered variable is smaller or equal a threshold value $c$. In case of a qualitative variable that decision is taken, if the considered variable is element of a well-chosen subset of the outcomes of that variable. At the so-called terminal node of such a tree one gets as result a forecast for the dependent variable, in our case for the expected value of the variable survival time. The above-mentioned threshold value is chosen in such a manner that the estimated variance within the nodes induced by that threshold is minimized – this for each independent variable. The construction of such a regression tree leads to a subset of independent variables – namely those variables, which give the best possible forecast of the dependent variable in the sense of variance.

Conclusion

The major issue of this study was to demonstrate the challenge when using several clinical, tissue and serum markers for the formulation of a comprehensive risk score. Due to the limited number of patients we did not make any further calculations to optimize any algorithm for predicting outcome. In addition to further identification of predictive markers we have to improve our capabilities to understand and interpret the data of informational networks. Facing the fact that from our patients we shall never have a complete data set of all possible impact factors, development of decision trees seems to offer promising options to overcome the challenge of predictive markers by formation of a collaborative network.

Acknowledgements

We gratefully thank A. Fiebeler and R. Schwab for their support in this study and their critical discussions helping us to interpret the large amount of data.
References

1 Billroth T: Chirurgische Erfahrungen, Zürich 1860-1867

2 Breiman L, Friedman JH, Olshen RA and Stone CJ: Classification and Regression Trees: Belmont, CA, Wadsworth

3 Chen DT, Hsu YL, Fulp WJ, Coppola D, Haura EB, Yeatman TJ and
Cress WD: Prognostic and predictive value of a malignancy-risk
gene signature in early-stage non-small cell lung cancer. J Natl

4 Dai J, Hu Z, Dong J, Xu L, Pan S, Jiang Y, Jin G, Chen Y and
Shen H: Host immune gene polymorphisms were associated with
the prognosis of non-small-cell lung cancer in Chinese. Int J

5 de Fraipont F, Levallet G, Creveuil C, Bergot E, Beau-Faller M,
Morin F, Milleron B and Zalcman G: An apoptosis methylation
prognostic signature for early lung cancer in the IFCT-0002 trial. Clin Cancer Res 18: 2976-2986,
2012.

6 Dehing-Oberije C, Aerts H, Yu S, De Ruyscher D, Menheere P,
Hilvo M, van der Weide H, Rao B and Lambin P: Development
and validation of a prognostic model using blood biomarker
information for prediction of survival of non-small-cell lung cancer
treated with combined chemotherapy and radiation or radiotherapy alone (NCT00181519, NCT00573040, and

7 Di Maio M, Krazkowski M, Fougerey R, Kowalski DM and
Gridelli C: Prognostic score for second-line chemotherapy of
advanced non-small-cell lung cancer: external validation in a
phase III trial comparing vinflunine with docetaxel. Lung Cancer

8 Geman S, Bienenstock E and Doursat R: Neural networks and

9 Giggerenzer G and Brighton H: Homo Heuristicus: Why Biased
Minds Make Better Inferences. Topics in Cognitive Science 1:
107-143, 2009.

10 Grenander U: On empirical spectral analysis of stochastic

11 Hu J, Liu YL, Piao SL, Yang DD, Yang YM and Cai L: Expression patterns of USP22 and potential targets BMI-1,
PTEN, p-AKT in non-small-cell lung cancer. Lung Cancer 77:
593-599, 2012.

12 Kossler W, Fiebeler A, Willms A, ElAidi T, Klosterhalfen B and
Klinge U: Formation of translational risk score based on
correlation coefficients as an alternative to Cox regression
models for predicting outcome in patients with NSCLC. Theor

13 Kuo YW, Jerng JS, Shih JY, Chen KY, Yu CJ and Yang PC: The
prognostic value of the simplified comorbidity score in the
treatment of small cell lung carcinoma. J Thorac Oncol 6: 3783-
381, 2011.

14 Liu H, Zhang T, Ye J, Li H, Huang J, Li X, Wu B, Huang X and
Hou J: Tumor-infiltrating lymphocytes predict response to
chemotherapy in patients with advance non-small cell lung

15 Lopez-Encuentra A, Lopez-Rios F, Conde E, Garcia-Lujan R,
Suarez-Gauthier A, Manes N, Renedo G, Duque-Medina JL,
Garcia-Lagarto E, Rami-Porta R, Gonzalez-Pont G, Astudillo-
Pombo J, Mate-Sanz JL, Freixinet J, Romero-Savedra T,
Sanchez-Cesperdes M and Gomez de la Camara A: Composite
anatomical-clinical-molecular prognostic model in non-small cell

172, 1981.

17 Navarro A, Diaz T, Gallardo E, Vinolas N, Marrades RM, Gel
B, Campayo M, Quera A, Bandres E, Garcia-Foncillas J,
Ramirez J and Monzo M: Prognostic implications of miR-16
expression levels in resected non-small-cell lung cancer. J Surg

18 Oken MM, Creech RH, Toracey DC, Horton J, Davis TE,
McFadden ET and Carbone PP: Toxicity and response criteria of
the Eastern Cooperative Oncology Group. Am J Clin Oncol 5:

19 Remmele W, Hildebrand U, Hienz HA, Klein PJ, Vierbuchten M,
Behnken LJ, Heicke B and Scheidt E: Comparative histological,
histochemical, immunohistochemical and biochemical studies on
oestrogen receptors, lectin receptors, and Barr bodies in human
breast cancer. Virchows Arch A Pathol Anat Histopathol 409:
127-147, 1986.


21 Wan YY, Beer DG and Guo NL: Signaling pathway-based
identification of extensive prognostic gene signatures for lung

22 Wu DW, Cheng YW, Wang J, Chen CY and Lee H: Paxillin
predicts survival and relapse in non-small cell lung cancer by

23 Zhang H and Singer BH: Recursive Partitioning and