

Reaching the limits of prognostication in non-small cell lung cancer: an optimized biomarker panel fails to outperform clinical parameters

Marianna Grinberg^{1,10}, Dijana Djureinovic^{2,10}, Hans RR Brunnström³, Johanna SM Mattsson², Karolina Edlund⁴, Jan G Hengstler⁴, Linnea La Fleur², Simon Ekman², Hirsh Koyi^{5,6}, Eva Branden^{5,6}, Elisabeth Ståhle⁷, Karin Jirström³, Derek K Tracy^{8,9}, Fredrik Pontén², Johan Botling², Jörg Rahnenführer^{1,11} and Patrick Micke^{2,11}

¹Department of Statistics, TU Dortmund University, Dortmund, Germany; ²Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden; ³Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, Lund, Sweden; ⁴Leibniz Research Centre for Working Environment and Human Factors (IfADo), TU Dortmund University, Dortmund, Germany; ⁵Department of Respiratory Medicine, Gävle Hospital, Gävle, Sweden; ⁶Centre for Research and Development, Uppsala University, County Council of Gävleborg, Gävle, Sweden; ⁷Department of Clinical Sciences, Uppsala University, Uppsala, Sweden; ⁸Oxleas NHS Foundation Trust, London, UK and ⁹Cognition, Schizophrenia and Imaging Laboratory, Institute of Psychiatry, Psychology and Neuroscience, King's College, London, UK

Numerous protein biomarkers have been analyzed to improve prognostication in non-small cell lung cancer, but have not yet demonstrated sufficient value to be introduced into clinical practice. Here, we aimed to develop and validate a prognostic model for surgically resected non-small cell lung cancer. A biomarker panel was selected based on (1) prognostic association in published literature, (2) prognostic association in gene expression data sets, (3) availability of reliable antibodies, and (4) representation of diverse biological processes. The five selected proteins (MKI67, EZH2, SLC2A1, CADM1, and NKX2-1 alias TTF1) were analyzed by immunohistochemistry on tissue microarrays including tissue from 326 non-small cell lung cancer patients. One score was obtained for each tumor and each protein. The scores were combined, with or without the inclusion of clinical parameters, and the best prognostic model was defined according to the corresponding concordance index (C-index). The best-performing model was subsequently validated in an independent cohort consisting of tissue from 345 non-small cell lung cancer patients. The model based only on protein expression did not perform better compared to clinicopathological parameters, whereas combining protein expression with clinicopathological data resulted in a slightly better prognostic performance (C-index: all non-small cell lung cancer 0.63 vs 0.64; adenocarcinoma: 0.66 vs 0.70, squamous cell carcinoma: 0.57 vs 0.56). However, this modest effect did not translate into a significantly improved accuracy of survival prediction. The combination of a prognostic biomarker panel with clinicopathological parameters did not improve survival prediction in non-small cell lung cancer, questioning the potential of immunohistochemistry-based assessment of protein biomarkers for prognostication in clinical practice.

Modern Pathology (2017) 30, 964–977; doi:10.1038/modpathol.2017.14; published online 10 March 2017

Patients with localized non-small cell lung cancer are potentially curable by surgical resection, but the risk of recurrence is high. (1) Adjuvant

chemotherapy has been proven to have a significant, but limited, effect, improving survival at 5 years by only 4%. (2) Thus, prognostication, or clinical stratification, is of particular relevance for this patient group. Patients with a low risk of relapse could, if accurately identified, be spared from adjuvant treatment. In contrast, patients with a poor prognosis might be expected to benefit from chemotherapy or other treatment modalities with novel compounds. This information might also help

Correspondence: Dr P Micke, MD, PhD, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala 751 85, Sweden. E-mail patrick.micke@igp.uu.se

¹⁰Shared first authorship.

¹¹Shared last authorship.

Received 9 December 2016; revised 6 January 2017; accepted 6 January 2017; published online 10 March 2017

patients to make informed choices about potential modalities of care.

In clinical practice, tumor stage, performance status, and age are the best predictors of overall survival and are used to guide therapy.³ However, as different outcomes are frequently observed for patients with similar clinicopathological characteristics, these factors are not sufficient. Consequently, much effort has been invested to identify better prognostic markers and various approaches have been applied. Genomic, transcriptomic, and proteomic studies of tumor tissue have led to the identification of numerous potential prognostic factors.^{4–8} Candidate protein biomarkers have been extensively evaluated using immunohistochemistry, which has the advantages of being cost-efficient and clinically feasible, as it is easily applicable on diagnostic formalin-fixed paraffin-embedded tissue. Indeed, numerous immunohistochemical studies have suggested a prognostic relevance for various proteins as single markers.^{5,9,10} Nevertheless, because of the relatively low prognostic impact and inconsistency in independent patient cohorts, no biomarker has been introduced in clinical diagnostics.^{5,9}

In contrast to gene expression signatures,^{11–13} only a few studies in non-small cell lung cancer have combined multiple protein biomarkers into one classifier, with the aims of increasing the prognostic power and of generating a robust and reproducible assay.^{14,15} In studies that have applied this strategy^{14,15} the proposed biomarkers were, however, not subsequently sufficiently validated to prove their value over traditional prognostic parameters. Important limitations in these lines of work included the statistical designs not adjusting for multiple testing and cutpoint optimization without validation in independent cohorts. Furthermore, potential markers were only compared to a selection of clinical parameters and not necessarily to an optimal combination.^{14,15} Finally, a combination of protein biomarkers with similar biological functions is likely to contain redundant prognostic information, reducing the likelihood of leading to an improved classification.

Our study intended to address these problems. First we selected a set of proteins with diverse biological functions and, in a next step, we constructed an optimized prognostic model using a large, clinically well-annotated non-small cell lung cancer patient cohort. The best-performing model was then applied to an independent validation cohort and compared with the most important clinical parameters. The stringent statistical design, the quality of the immunohistochemical annotation, and the completeness of both non-small cell lung cancer patient cohorts make this study unique, and we believe that it provides a realistic estimation of the prognostic potential of protein biomarkers in non-small cell lung cancer.

Materials and methods

Patient Cohorts and Clinical Characteristics

The study material comprised two patient cohorts with primary non-small cell lung cancer, surgically treated at the University Hospital in Uppsala, Sweden. Uppsala cohort I included 354 non-small cell lung cancer patients treated in 1995–2005,^{4,16} and Uppsala cohort II included 357 patients treated in 2006–2010.^{4,17,18} Formalin-fixed paraffin-embedded tissue from both cohorts was used to construct tissue microarrays for the immunohistochemical analysis. The clinical characteristics (age at diagnosis, gender, smoking history, performance status according to World Health Organisation (WHO) criteria, tumor stage (TNM 7th edition),¹⁹ and tumor histology in accordance with the WHO classification of 2004 (ref. 20) of the patients that were included in the final analysis are shown in Supplementary Table 1. The study was performed in accordance with the Swedish Biobank Legislation and was approved by the Uppsala University Ethical Review (Reference 2006/325, Uppsala cohort I; Reference 2012/532, Uppsala cohort II).

Selection of Protein Biomarkers

For selection of the biomarker panel, a pipeline was applied based on following criteria: (1) A systematic evaluation of protein markers reported in the scientific literature between 2008 and 2013,⁵ consistent prognostic association in at least two studies and consistent results in at least 50% of the studies.⁵ (2) Prognostic significance (adjusted *P*-value < 0.05) for at least one probe set in a meta-analysis based on Affymetrix gene expression data from 10 independent cohorts, comprising in total 1779 non-small cell lung cancer patients; 1142 adenocarcinomas, 451 squamous cell carcinomas, and 186 other non-small cell lung cancer histologies (Supplementary Table 2A). (3) Availability of a reliable antibody in the Human Protein Atlas database (www.proteinatlas.org). Antibodies were chosen if the staining pattern was in accordance with the expected subcellular and histological expression in the scientific literature (Supplementary Table 2B). (4) Involvement in different tumorigenic mechanisms was based on information in UniProt and corresponding literature (Supplementary Table 2B). As an additional biomarker, we included cell adhesion molecule 1 (CADM1), which fulfilled all criteria except that it was reported before 2008.^{4,21} The results from the meta-analysis are shown in Supplementary Figure 1. The selection procedure is illustrated in Figure 1.

Meta-Analysis

The meta-analysis was performed as previously described,⁴ including 10 gene expression array data

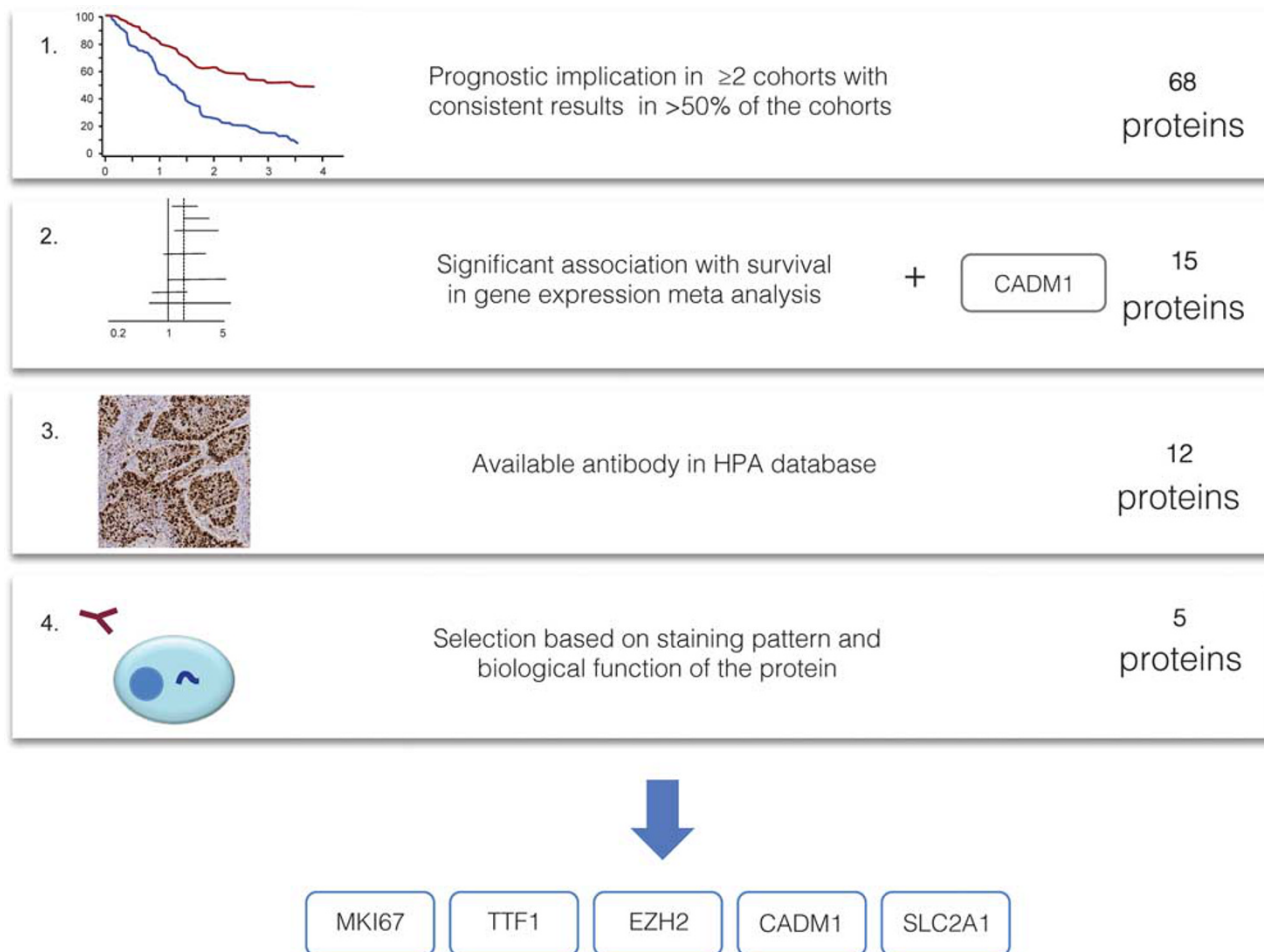


Figure 1 Pipeline of protein selection. For selection of the biomarker panel, a pipeline was applied based on subsequent criteria: (1) prognostic implication in the scientific literature with consistent prognostic results in at least half of at least two independent cohorts, reported between 2008 and 2013. (2) Prognostic significance (adjusted P -value < 0.05) of at least one probe set in the gene expression meta-analysis of all non-small cell lung cancer histologies and either adenocarcinoma or squamous cell carcinoma patients. As an additional biomarker we included cell adhesion molecule 1 (CADM1) that did not fulfill criterion 1 (reported before 2008). (3) Availability of a reliable antibody in the Human Protein Atlas database. (4) Involvement in different biological mechanisms. The genes and corresponding information included in the selection process are listed in Supplementary Table 2.

sets based on Affymetrix microarrays (GSE37745,⁴ GSE14814,¹³ GSE19188,²² GSE29013,²³ GSE30219,²⁴ GSE31210,²⁵ GSE3141,²⁶ GSE4573,²⁷ GSE50081,²⁸ and Shedden *et al.* 2008 (ref. 11)). Meta-analysis was performed with random effect models. Results were visualized with forest plots, and significance of the overall effect was measured with the P -value of the random effect models. All P -values were two-sided and adjusted for multiple testing for all 54675 analyzed probe sets with the Benjamini–Hochberg procedure.²⁹ The meta-analysis was conducted using the R package ‘meta’ (<http://CRAN.R-project.org/package=meta>).

Tissue Microarray Production and Immunohistochemistry

The selected proteins (MKI67, TTF1, EZH2 (enhancer of zeste homolog 2), CADM1, and SLC2A1) were

stained and analyzed in both the Uppsala cohort I and Uppsala cohort II. Tissue microarray construction and immunohistochemistry were performed as previously described.³⁰ In brief, representative formalin-fixed paraffin-embedded tissue from donor blocks were punched (1 mm in diameter) using a manual tissue arrayer (MTA-1, Beecher Instruments Sun Prairie, WI, USA) and placed in a recipient block, generating tissue microarrays containing tissues in total from 711 non-small cell lung cancer patients (354 from cohort I and 357 from cohort II) represented in duplicates. Sections of 4 μm of the tissue microarray blocks were cut using a microtome (HM 355S, Microm), mounted on adhesive slides (SuperFrost Plus, Thermo Scientific, Braunschweig, Germany), and baked for 45 min at 60 °C. Deparaffinization and hydration were performed in xylene and graded alcohols to distilled water prior to immunohistochemical staining. Blocking for

endogenous peroxidase was done using 0.3% hydrogen peroxide in 95% ethanol for 5 min. For antigen retrieval, a pressure boiler (Decloaking chamber, Biocare Medical, Walnut Creek, CA, USA) was used and the slides were boiled for 4 min at 125 °C in citrate buffer, pH6 (Lab Vision, Fremont, CA, USA). Automated immunohistochemistry was performed using an Autostainer 480 instrument (Thermo Fisher Scientific, Runcorn, UK). Primary antibodies used for immunohistochemical analysis included the following: CAB000058, DakoCytomation, clone MIB, dilution 1:200, targeting MKI67; CAB000078, DakoCytomation, clone 8G7G3/1, dilution 1:150, targeting TTF1; CAB009589, Novocastra, clone 6A10, dilution 1:500, targeting EZH2; CAB037266, Sigma, polyclonal antibody, dilution 1:10 000 targeting CADM1 and HPA058494, Atlas Antibodies, polyclonal antibody, dilution 1:50 targeting SLC2A1. The tissue microarrays were incubated with primary antibodies diluted in UltraAb Diluent (Lab Vision) and the secondary reagent UltraVision LP HRP polymer (Lab Vision) for 30 min each at room temperature. Following washing steps, the slides were developed for 10 min at room temperature, adding diaminobenzidine (Lab Vision) as a chromogen, and thereafter counterstained with Mayer's hematoxylin (Histolab, Gothenburg, Sweden) and mounted with Pertex (Histolab). The stained slides were scanned at $\times 20$ magnification using an Aperio ScanScope XT Slide Scanner (Aperio Technologies, Vista, CA, USA) to obtain high-resolution digital images for the annotation of protein expression. An immunohistochemistry score was calculated by multiplying the staining intensity (negative=0, weak=1, moderate=2, and strong=3) with the fraction of stained tumor cells (1=0–1%, 2=2–10%, 3=11–20%, 4=21–30%, 5=31–40%, 6=41–50%, 7=51–75%, and 8 \geq 75%), giving a range of 0–24. This immunohistochemistry score was used for further analyses.

Statistical Analysis

Survival analysis. All analyses were performed using the statistical programming language 'R-version 3.1.1'. Overall survival was calculated from the date of diagnosis to the date of death. The survival times were censored at 5 years. Survival was analyzed by univariate and multivariate Cox models and visualized by Kaplan–Meier plots. Survival functions were compared with the log-rank test using the R package 'survival'.³¹ The Kaplan–Meier plots were generated based on dichotomized immunohistochemistry and risk scores (see below 'Best prognostic model' and 'Assessment of model performance' for definitions of selected cutpoints). The clinicopathological variables with an established prognostic association—tumor stage, performance status, and age at diagnosis—were categorized as follows for all analyses: stage I vs stage II–IV,

performance status 0 vs performance status I–IV, ≤ 70 vs > 70 years. Multivariate Cox analyses were performed with inclusion of the above-mentioned clinicopathological variables, together with all possible combinations of the immunohistochemistry scores based on each protein's best cutpoint (see below 'Best prognostic model') to assess the prognostic power of each combined model. The prognostic power of each model was assessed by the concordance index (see below 'C-index'). Adjustment for multiple testing was done by the Benjamini–Hochberg method.²⁹

C-index. The C-index is a rank-based method for assessing the prognostic power of a model³² and was here applied to indicate how well a model discriminated patients with longer survival from patients with shorter survival times. On the basis of a fitted Cox model, the C-index compares the predicted survival times with the observed survival times of all possible patient pairs, and estimates the probability of concordant patient pairs. A patient pair is concordant if the predicted outcomes agree with the actual outcomes, ie, if the predicted survival time is longer for the patient who lived longer. Thus, a patient pair is only informative if the patient with shorter survival time has died, and only the patient pairs that fulfilled this criterion were included in the analysis. A C-index of 1 implies perfect prediction accuracy, a C-index of 0.5 indicates no predictive ability, and a value below 0.5 indicates a predictive ability that is even worse than random guessing.

Best prognostic model. The predictive power of each individual protein in combination with the dichotomized clinicopathological variables was first assessed. To accomplish this, multivariate Cox models, based on dichotomized immunohistochemistry scores, were fitted in the Uppsala cohort I. For each protein, different cutpoints were considered by splitting the data into two groups—below and above the cutpoint—at each possible protein score (range 0–24). A multivariate Cox model was fitted for each split, and the corresponding C-index was calculated. For each protein, the cutpoint corresponding to the model with the highest C-index was selected. This resulted in five fixed cutpoints, referred to as the five proteins' best cutpoints, which were used in all subsequent analyses.

Next, we aimed to define the best prognostic model, based on the clinicopathological variables together with an optimal combination of the protein scores, using the above-defined best cutpoint for each protein. To this end, we fitted multivariate Cox models that included the dichotomized clinicopathological variables together with all possible combinations of two to five proteins, followed by C-index calculation. The best-performing model was defined as the model that yielded the highest C-index.

Finally, a risk score was calculated for each individual patient, where a higher risk score meant a higher risk of death. On the basis of the best prognostic model, the risk score of a patient was defined as the linear combination of the fitted parameters and the patient's individual values for the fitted parameters (ie, the immunohistochemistry scores dichotomized according to each protein's best cutpoint and the clinicopathological variables dichotomized at the above-described fixed cutoffs). Given the best prognostic model based on the dichotomized variables (clinicopathological and

immunohistochemistry scores), the risk score for a patient was calculated as follows:

$$\begin{aligned} \text{risk score} = & \text{age} \cdot x^{\text{age}} + \text{stage} \cdot x^{\text{stage}} \\ & + \text{performance status} \cdot x^{\text{performance status}} \\ & + \sum_{k=1}^n \text{protein}^k \cdot x^k, \end{aligned}$$

where age, stage, and performance status denote the estimated coefficients of the clinicopathological variables and protein^k the estimated coefficient of the k^{th} -protein, $k \in \{\text{MKi67, EZH2, TTF1, SLC2A1, CADM1}\}$, obtained from the fitted model, and

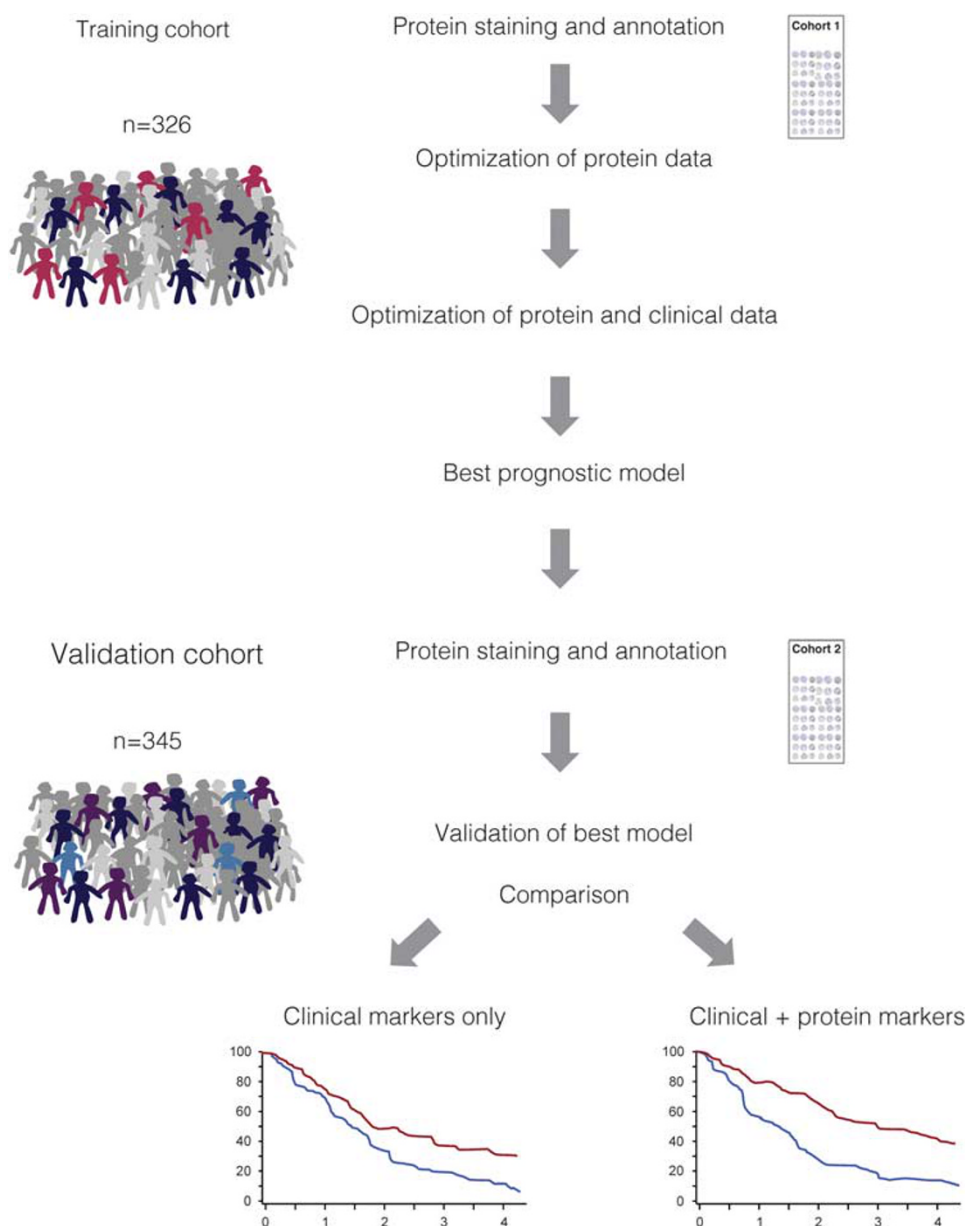


Figure 2 Study design. Tissue microarrays of the training cohort (Uppsala cohort I, $n = 326$) were annotated and immunohistochemistry scores were obtained for all five proteins. The different immunohistochemistry scores were used to identify the best cutoff to predict survival. The fixed combinations for clinicopathological parameters (age, stage, and performance status) with all possible combinations of protein markers were tested to develop the best prognostic model. The best prognostic model obtained on the training cohort was tested in the validation cohort (Uppsala cohort II, $n = 345$) and compared to the model only consisting of clinicopathological parameters.

x^{variable} , variable ϵ {age, stage, performance status, k }, denotes the indicated individual value for patient x . A risk score was also calculated for each patient based on the clinicopathological data only.

Assessment of model performance. In the next step, we evaluated the performance of the best-performing model with regard to prediction of overall survival rates in an independent validation cohort (Uppsala cohort II), and compared the best-performing model, based on protein and clinicopathological data, with models based on clinicopathological or protein data only. This was accomplished in two ways. First, we calculated the C-index. Second, we calculated the *sensitivity* of the model as the rate of patients with high risk scores among the short-time survivors, and the *specificity* as the rate of patients with low risk score among the long-time survivors based on a 2×2 contingency table of dichotomized survival times and risk scores. For survival time, the cutpoints for dichotomization were 2, 3, and 4 years. For the risk score, the cutpoint was chosen so that the proportion of patients with high risk scores equaled the proportion of patients with survival times shorter than 2, 3, and 4 years. For the direct comparison of two models we first calculated for each patient if survival time and risk score agreed (correct prediction: if survival is long and risk score is low, or if survival is short and risk score is high), and then compared the predictions of the two models (correct, false) in 2×2 contingency tables. To assess the statistical significance of the difference of two models we applied the McNemar's test to the contingency tables. A small two-sided P -value ($P \leq 0.05$) indicates that one model makes more correct predictions than the other model.

Receiver-operating-characteristic curves were used to visualize the relationship between survival time (dichotomized at 4 years) and risk score (continuous). The patients who died within the first 4 years were labeled as positives, and those who lived beyond 4 years as negatives, ie, the patients with a high risk score who died before 4 years were labeled as true positives, and those with a low risk score who lived beyond 4 years were classified as true negatives. The true-positive rate was plotted against the false-positive rate, which is equal to $1 - \text{specificity}$, in the receiver-operating-characteristic curve.

Results

Selection of Clinical and Protein Markers for the Prognostic Panel

The study design, based on the training and validation cohorts, is illustrated in Figure 2. The three clinicopathological parameters (stage, age, and performance status) analyzed in this study have a well-established prognostic value and are those most commonly used to stratify patients for standard

treatment or in clinical trials. As expected, all three were associated with overall survival in the training cohort (Uppsala cohort I; Supplementary Figure 2). The selection process for the prognostic panel, illustrated in Figure 1, identified five proteins with different tumorigenic mechanisms:

Antigen Ki-67 (MKI67) is expressed during the active phases of the cell cycle (G1, G2, and S) and serves as a marker of proliferation.³³ While in breast cancer and neuroendocrine tumors MKI67 is an established prognostic and diagnostic marker,^{34,35} the use of MKI67 in lung cancer is not established, although its potential prognostic value has been demonstrated in several studies.³⁶

Homeobox protein Nkx-2.1 (NKX2-1), also known as TTF1 (thyroid transcription factor-1), is a transcription factor, exclusively expressed in thyroid, lung, and ventral forebrain. In the lung, TTF1 is involved in morphogenesis and differentiation of epithelial cells.³⁷ TTF1 has an established role in tumor development and is a diagnostic marker for the origin of cancer and the adenocarcinoma differentiation.³⁸ Several studies indicate that higher TTF1 expression is associated with a better prognosis.^{39,40}

The enhancer of zeste homolog 2 (EZH2) is the functional unit of the polycomb repressive complex 2, a methyltransferase that mediates gene silencing through post-translational histone modifications, and works in principal as a transcriptional repressor.⁴¹ High expression of EZH2 has been reported in a wide range of cancers and higher expression has been linked to more aggressive tumor behavior.⁴²⁻⁴⁴

The CADM1 belongs to the immunoglobulin superfamily and is involved in cell adhesion, proliferation, and differentiation.⁴⁵ CADM1 acts as a tumor suppressor in several epithelial cancers and lower expression of CADM1 has been associated with worse prognosis in epithelial cancers, including lung cancer.^{46,47}

Solute carrier family 2, facilitated glucose transporter member 1 (SLC2A1 alias GLUT1), is a transporter protein involved in cellular glucose metabolism.⁴⁸ Overexpression of SLC2A1 is reported in several cancers and has also been associated with poorer survival in lung cancer.⁴⁹

Annotation of Protein Expression and Cutpoint Optimization

The five protein biomarkers were analyzed with immunohistochemistry on the Uppsala cohort I tissue microarray, including 326 evaluable tumors. Representative staining patterns and the distribution of the protein scores are shown in Figure 3. To identify the cutpoints of the protein scores that best discriminated between long- and short-term survivors, the C-index was used as a measurement of prognostic performance (Figure 4 and Supplementary Figure 3). For each protein (MKI67, TTF1, EZH2, CADM1, and SLC2A1), the training set was

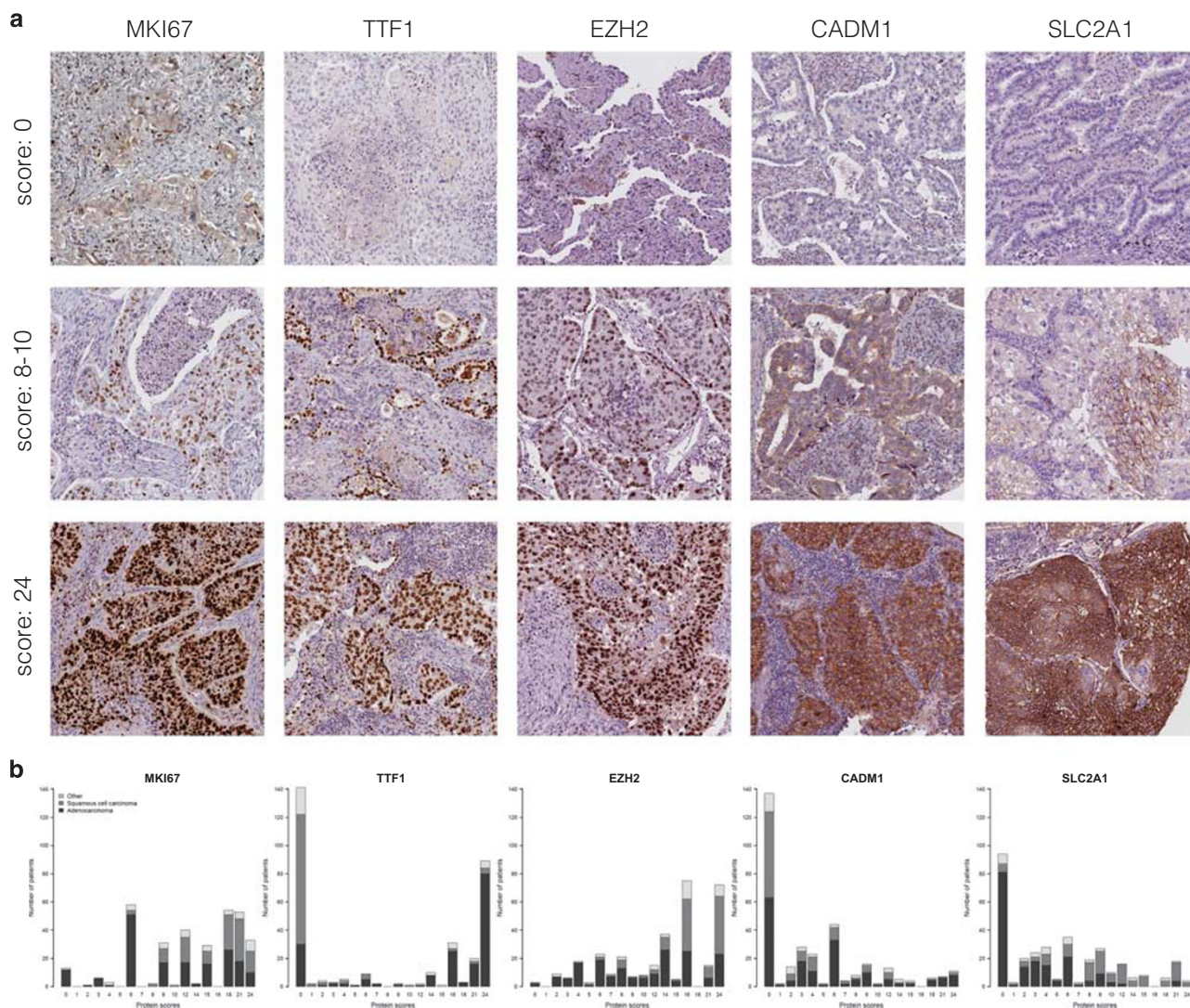


Figure 3 (a) Staining patterns of the selected proteins. Representative immunohistochemical images of the five proteins in non-small cell lung cancer. The staining intensity and fraction of positive cells were annotated. The product of both resulted in an immunohistochemistry score that was used in further analysis. (b) Bar plots showing the distribution of the immunohistochemistry scores.

split into two groups at each possible protein score (range 0–24) and for each split both univariate and multivariate model (including age, stage, and performance status) were fitted, followed by calculation of the C-index. The analysis was performed separately for all non-small cell lung cancer, adenocarcinomas, and squamous cell carcinomas; this procedure was repeated for each protein (MKI67, TTF1, EZH2, CADM1, and SLC2A1). In the final prognostic model, the cutpoint with the highest C-index based on the multivariate analysis was used for dichotomization of the protein scores (Supplementary Table 3).

Development of best-Performing Prognostic Model Based on Clinical and Protein Data

On the basis of the protein's best cutpoints, we first performed univariate and multivariate Cox regression models to analyze the association of each

protein with overall survival, alone (Supplementary Table 4) and combined with the clinical data (Supplementary Table 5). All proteins showed a significant, or near significant, association with overall survival, either in the complete non-small cell lung cancer cohort or in the separate analysis of the adenocarcinomas, with C-index values ranging from 0.54 to 0.58 (Supplementary Table 6). The results were illustrated using Kaplan–Meier plots (Figure 4 and Supplementary Figure 3). Compared with the clinical parameters, the single protein markers showed comparable associations with overall survival (Supplementary Table 6).

Next, the analysis was repeated for the combination of the five proteins. This improved the C-index for the complete cohort (0.59), and for the histological subtypes (adenocarcinoma: 0.63; squamous cell carcinoma: 0.58), compared to the C-indices obtained when the proteins were analyzed

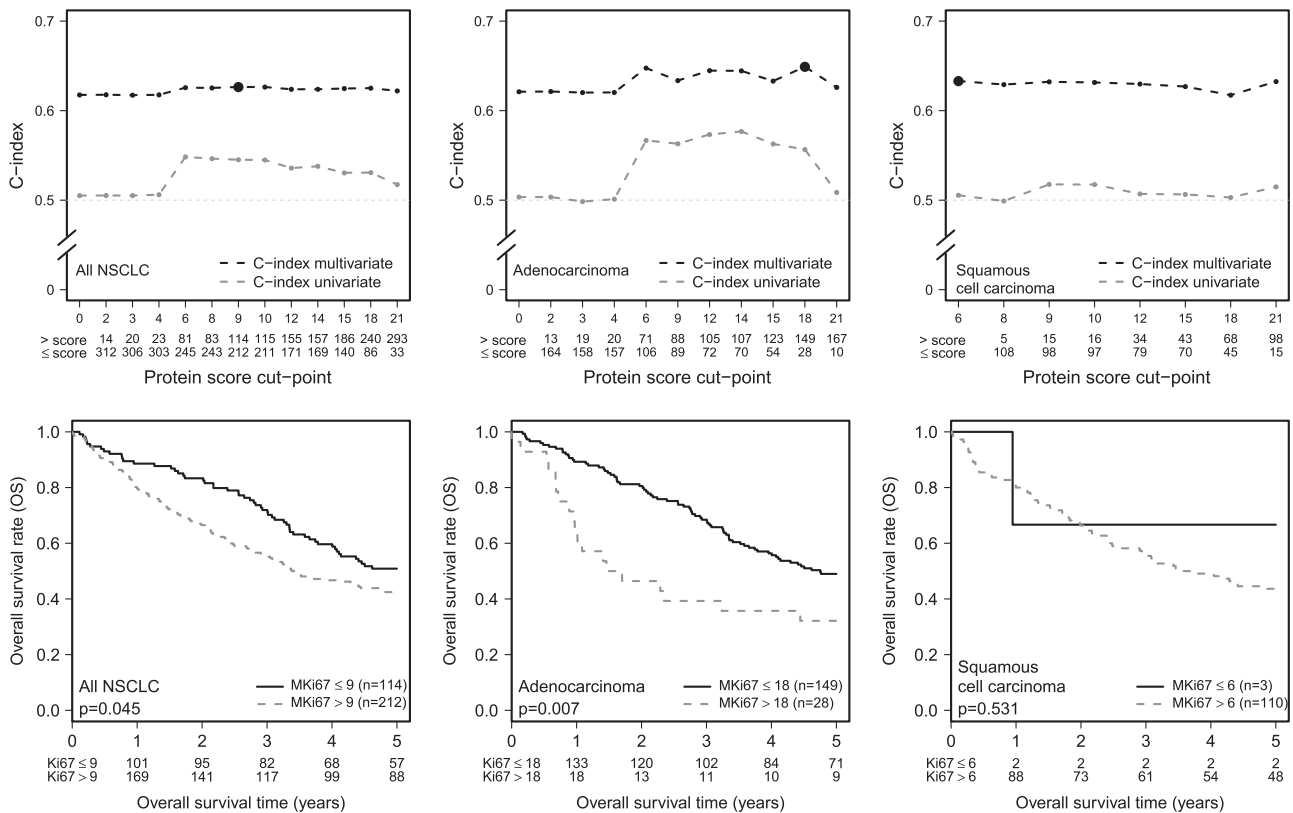


Figure 4 Identification of best cutoffs for immunohistochemistry scores. Plot of the C-indices obtained from the trained uni- and multivariate Cox models using the dichotomized immunohistochemistry scores for MKI67 for all non-small cell lung cancer cases, adenocarcinoma, and squamous cell carcinoma. The univariate model (gray dashed line) was built based on the protein alone, the multivariate model (black dashed line) combines clinical data (age, stage, performance status) with single protein data. The C-index (y axis) was calculated for the univariate (gray dotted line) and multivariate model (black dotted line) using all possible protein cutpoints (x axis). The best cutpoint for dichotomizing the protein score was determined by the highest multivariate C-index (black bold dot). The light gray dashed line indicates the boundary line to random guessing. Kaplan–Meier analysis of overall survival stratified by MKI67 protein expression dichotomized at its best cutpoint. The corresponding plots for the other four proteins are shown in Supplementary Figure 3.

separately. However, the C-index was not higher than that obtained by a combination of the clinical parameters only (all non-small cell lung cancer: 0.62; adenocarcinoma: 0.62; squamous cell carcinoma: 0.63; Supplementary Table 6).

Finally, the best prognostic model was defined based on the clinicopathological variables together with an optimal combination of the protein scores, yielding the highest C-indices (all non-small cell lung cancer: 0.64; adenocarcinoma: 0.69; squamous cell carcinoma: 0.66; Supplementary Table 6). The best model for all non-small cell lung cancer included the clinical parameters combined with all five proteins. For the adenocarcinoma subgroup, the best model included the clinical parameters combined with MKI67, EZH2, TTF1, and CADM1, and for the squamous cell carcinomas it included the clinical parameters combined with EZH2, TTF1, SLC2A1, and CADM1 (Supplementary Table 7). The best models (highest C-index) are shown in Supplementary Table 8. Kaplan–Meier curves were plotted for the complete non-small cell lung cancer cohort, as well as for the two main histologies separately, with patients stratified at dichotomized

risk scores (Figure 5). The models were subsequently applied to the validation cohort.

Independent Validation of the Best-Performing Models

To validate the models that performed best in the training cohort for all non-small cell lung cancer, adenocarcinoma, and squamous cell carcinoma, we next applied them to an independent cohort (Uppsala cohort II) and compared them to the models based on only the clinicopathological variables and only the protein biomarkers.

In the validation cohort, the models consisting of only the clinicopathological variables revealed C-indices of 0.63 (all non-small cell lung cancer), 0.66 (adenocarcinoma), and 0.57 (squamous cell carcinoma), and the models based only on the five protein biomarkers demonstrated C-indices of 0.57 (all non-small cell lung cancer), 0.65 (adenocarcinoma), and 0.54 (squamous cell carcinoma).

In comparison, the previously established best-performing prognostic models, combining clinical parameters and an optimal combination of the

Training cohort

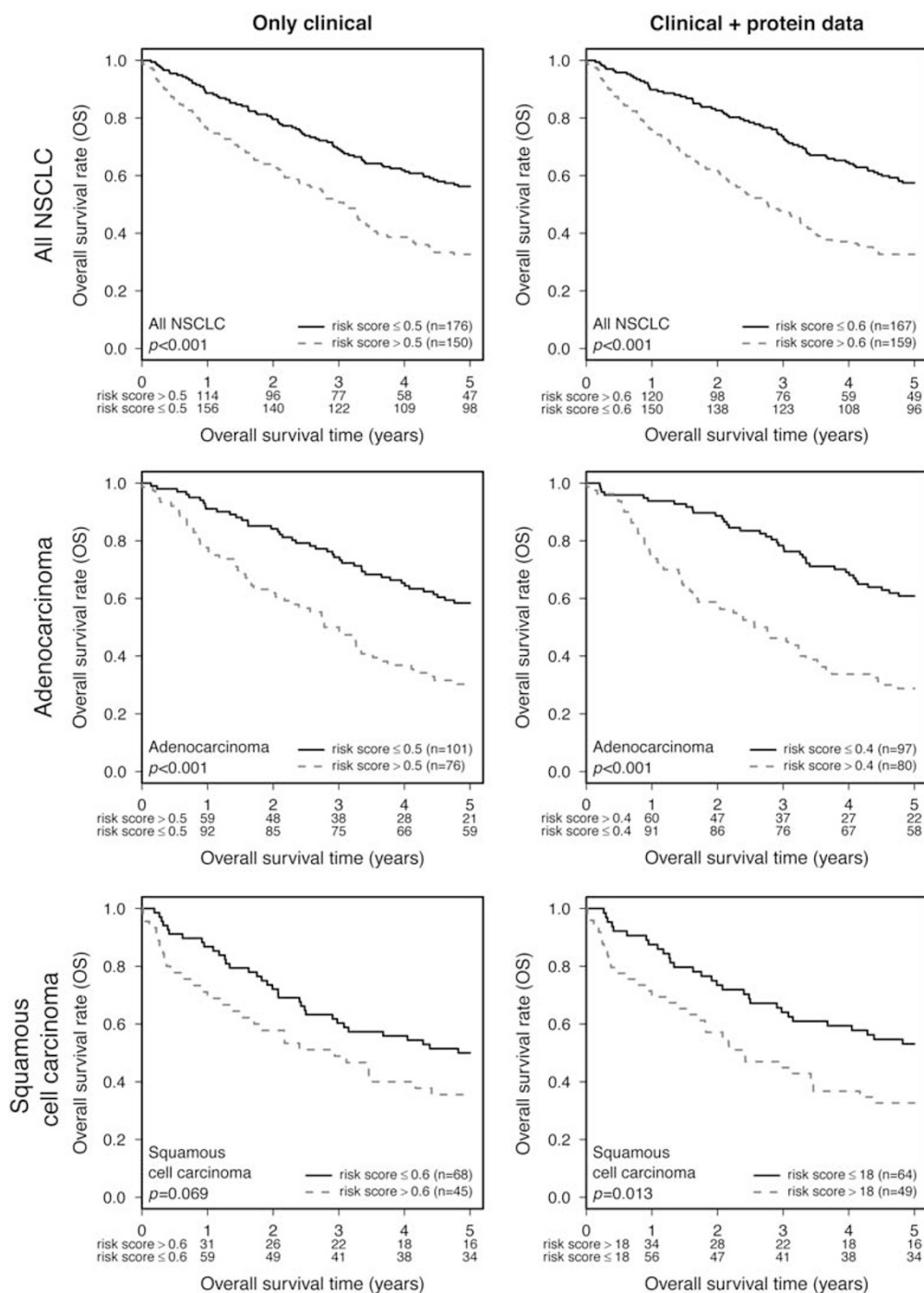


Figure 5 Overall survival of the patients from the training cohort stratified by the risk score. Risk scores were calculated from the model trained on the clinical variables only (age, stage, performance status; left column) and in combination with the protein data (right column), given the best-parameter combination with highest-trained C-index. The cutpoint for stratification of the risk score was chosen such that the proportion of patients with high risk scores equaled the proportion of patients with survival time shorter than 4 years. The unadjusted P -value of the log-rank test is given in the figure.

protein markers, revealed higher C-indices for the complete non-small cell lung cancer cohort (0.64) and adenocarcinomas (0.70), but not for the squamous cell carcinomas (0.56; Table 1).

Comparable results were obtained based on receiver-operating-characteristic curves when the clinical model consisting of the three clinical parameters only was compared with the best model.

Table 1 C-indices of the best models on the validation cohort using only clinical data or clinical data in combination with protein data

	Only clinical ^a	Clinical and protein ^b	Only proteins ^c
All NSCLC	0.632	0.641	0.568
Adenocarcinoma	0.660	0.701	0.646
Squamous cell carcinoma	0.565	0.562	0.539

Abbreviations: C-index, concordance index; NSCLC, non-small cell lung cancer.

Protein data dichotomized at best cutpoint.

Age at diagnosis: ≤ 70 years vs > 70 years; stage: IA+IB vs II–IV.

Performance status according to WHO 0 vs 1–4.

^aRefers to all three clinical parameters combined (age, stage, performance status according to WHO).

^bRefers to the best clinical + protein combination.

^cRefers to all five proteins.

The area under the curve was markedly higher for the combined model only when the adenocarcinoma cases were analyzed (0.71 vs 0.75, Supplementary Figure 4). The results of these analyses were illustrated using Kaplan–Meier plots (Figure 6).

Although the C-index gives an estimation of the model performance, the comparison above does not provide information whether or not the difference between two C-indices is statistically significant, such as the difference between 0.70 (best-performing model) and 0.66 (model based on clinical data only) observed in the validation cohort for the adenocarcinoma subgroup. To address this question, we first predicted if each individual patient survived longer than 2, 3, or 4 years, respectively, based on the best-performing model and model built based on clinical parameters only, and then compared the predicted outcome with the actual outcome of the patient (Supplementary Table 9). The combination of clinical and protein markers was *not* found to correctly classify a significantly higher number of patients as long- or short-term survivors beyond 2, 3, or 4 years (adjusted *P*-value > 0.08 , all comparisons).

Discussion

The choice of therapy for lung cancer patients is based on clinical parameters, most importantly stage, performance status, and age. All three parameters are associated with prognosis and are consequently used to guide therapy decisions. This prognostic accuracy is of particular importance for patients with localized disease, for whom surgery presents a potentially curative treatment option. Since most patients develop local or distant relapse, adjuvant therapy, with the aim to target remaining tumor cells, is added. However, the effect of this adjuvant intervention is modest, with improvement of 5-year survival rates by only 4%.² This means that only 1 of 25 patients benefits from this demanding therapy,

whereas 24 of 25 patients suffer from side effects without any benefit.

With this background, we developed and validated an immunohistochemistry-based biomarker assay that adds prognostic information to that conveyed by the most important clinical parameters. A protein biomarker panel was selected based on supportive information from the scientific literature, and validation of significant survival associations on the transcript level in a large collection of 10 publically available non-small cell lung cancer data sets (1779 patients). Furthermore, a stringent biostatistical approach was applied to be able to critically assess the prognostic value of the models. In the direct comparison, the prognostic model based on proteins alone failed to outperform clinical parameters. Combining the protein biomarkers with the clinical parameters demonstrated only limited added value, and would appear to be of minor relevance for clinical practice. It should also be noted that the performance of the biomarker immunohistochemistry assay is likely to be overestimated, since both the training and the validation cohort originated from the same center, were stained in the same laboratory, and were annotated by the same observer, ie, interlaboratory and interobserver variability, which might further impair the performance, were excluded.

So why did the combined prognostic model fail? Obviously, the choice of protein biomarkers can be questioned. Each of the five selected proteins showed a significant or close to significant prognostic impact in the training cohort, with hazard ratios between 0.6 and 0.8 for favorable prognostic markers (CADM1, TTF1) and 1.3 and 1.4 for unfavorable markers (MKI67, EZH2, SLC2A1), depending on histology. This was in line with previous studies evaluating these biomarkers^{21,39,43,50,51} and the hazard ratios were even higher than those obtained in the meta-analysis of publically available gene expression cohorts for the corresponding transcripts. Of note, the size of the hazard ratios was in the range of many other proposed biomarkers,^{14,15,52} with few exceptions.^{53,54} Thus, protein selection was most likely not decisive for the failure of the overall procedure. Upon first sight, the combination of the five proteins suggested an impressive separation of the survival curves both in the training and validation cohort (Figure 4 and Supplementary Figure 3). Nevertheless, this separation was not better than stratification solely based on the combination of clinical parameters. This result was already obtained in the training cohort, where the clinical parameters alone or in combinations showed higher C-indices and hazard ratios, ie, the combination of tumor-(stage) and patient-related factors (age and performance status) in general outperformed molecular tumor features. Only the addition of the protein markers increased the prognostic power, but if this minimal increase is of any practical relevance is questionable. Our study was not able to demonstrate a significantly improved prediction of 2, 3, or 4-year

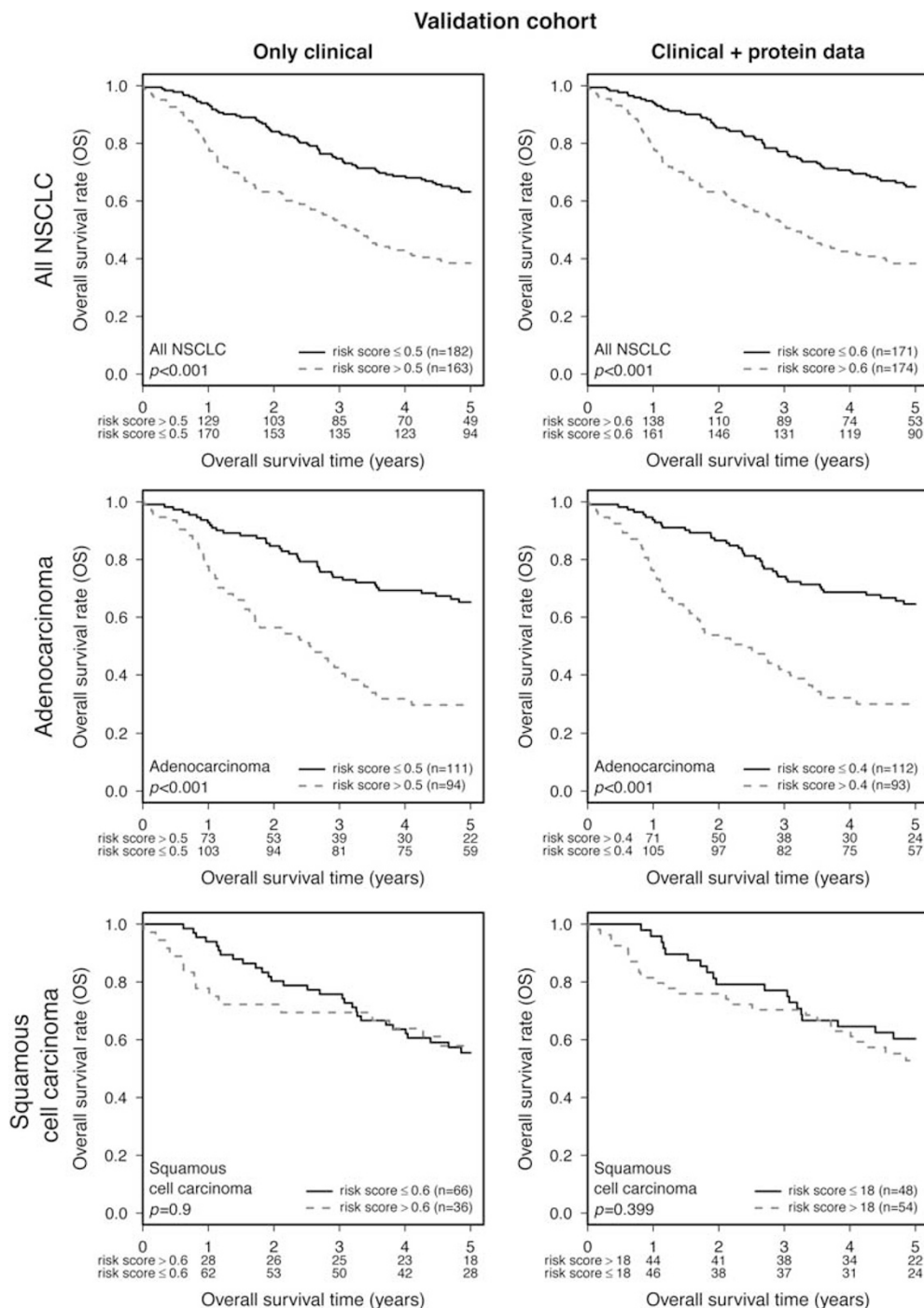


Figure 6 Overall survival of the patients from the validation cohort stratified by the risk score. Kaplan–Meier plots were established for all non-small cell lung cancer cases (upper), adenocarcinomas (middle), and squamous cell carcinoma (lower) separately. The best-performing model combining clinical and protein data based on the analysis of the training cohort was applied to the validation cohort (right) and compared to a model of clinical parameters only (left). The cutpoint for stratification of the risk score was adopted from the training cohort. The unadjusted P -value of the log-rank test is given in the figure.

survival for the individual patients. These findings obviously question the general concept that immunohistochemical markers have an additional value for prognostication in localized lung cancer.

Are there better methods for molecular prognostication? Perhaps a more promising strategy is the use of global gene expression profiles to develop

prognostic classifiers. The public availability of gene expression data sets facilitated validation across multiple independent patient cohorts and several of them showed promising, and stage-specific, performance.^{55,56} Two of them were adapted for the use of formalin-fixed paraffin-embedded tissue in a quantitative real-time PCR format and were

commercially launched to predict survival after radical resection.^{57,58} Although both assays demonstrated significant separation of patients with short- and long-term survival within stage I or even stage Ia patients, neither was tested head to head to clinical models including performance status. Thus, we believe that molecular prognostication is yet to provide proof that it can add substantial information regardless of whether protein or gene expression as biomarkers are used. In contrast, our study reconfirms the importance of traditional clinical parameters for prognostication. This should motivate clinicians to assess these parameters as accurately as possible to obtain optimal prognostic information. Attempts are ongoing to refine the TNM staging system for non-small cell lung cancer, and the assessment of patient performance status may also be an appropriate subject for optimization. The implementation of additional patient-related factors may further optimize survival prediction. Promising factors to be included in such an extended model include, for instance, pre-operative weight loss⁵⁹ and the Glasgow prognostic score based on plasma levels of C-reactive protein and albumin.⁶⁰

Finally, it should be stressed that reporting of the prognostic impact of a molecule is not superfluous. A significant survival association might, for instance, indicate a particular molecular tumor subgroup, eg, TTF1,¹⁶ or a tumorigenic mechanism (eg, EGFR,⁶¹ CADM1 (ref. 62)). Here, we presented a stringent statistical approach to develop and validate an immunohistochemical predictor of survival of non-small cell lung cancer after surgical resection. However, the failure to substantially improve prognostic accuracy, alone or together with clinical parameters, challenges efforts to implement immunohistochemistry-based assays for prognostication.

Acknowledgments

We acknowledge the Human Protein Atlas team for the support of tissue microarray construction and immunohistochemical staining. We thank the Clinical Pathology at the Uppsala University Hospital and Simin Tahmasebpoor for assistance with tissue samples and sample preparations. This study was supported by the Swedish Cancer Society and Lions Cancer Foundation, Uppsala, Sweden, and the Regional Research Foundation of the Uppsala Örebro region (RFR 556381).

Disclosure/conflict of interest

The authors declare no conflict of interest.

References

1 Rami-Porta R, Crowley JJ, Goldstraw P. The revised TNM staging system for lung cancer. *Ann Thorac Cardiovasc Surg* 2009;15:4–9.

- 2 Burdett S, Pignon JP, Tierney J, *et al*. Adjuvant chemotherapy for resected early-stage non-small cell lung cancer. *Cochrane Database Syst Rev* 2015;2: CD011430.
- 3 Sculier JP, Chansky K, Crowley JJ, *et al*. The impact of additional prognostic factors on survival and their relationship with the anatomical extent of disease expressed by the 6th Edition of the TNM Classification of Malignant Tumors and the proposals for the 7th Edition. *J Thorac Oncol* 2008;3:457–466.
- 4 Botling J, Edlund K, Lohr M, *et al*. Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin Cancer Res* 2013;19:194–204.
- 5 Lindskog C, Edlund K, Mattsson JS, *et al*. Immunohistochemistry-based prognostic biomarkers in NSCLC: novel findings on the road to clinical use? *Expert Rev Mol Diagn* 2015;15:471–490.
- 6 Micke P, Edlund K, Holmberg L, *et al*. Gene copy number aberrations are associated with survival in histologic subgroups of non-small cell lung cancer. *J Thorac Oncol* 2011;6:1833–1840.
- 7 Zhu CQ, Tsao MS. Prognostic markers in lung cancer: is it ready for prime time? *Transl Lung Cancer Res* 2014;3: 149–158.
- 8 Wen J, Fu J, Zhang W, *et al*. Genetic and epigenetic changes in lung carcinoma and their clinical implications. *Mod Pathol* 2011;24:932–943.
- 9 Zhu CQ, Shih W, Ling CH, *et al*. Immunohistochemical markers of prognosis in non-small cell lung cancer: a review and proposal for a multiphase approach to marker evaluation. *J Clin Pathol* 2006;59:790–800.
- 10 Kern SE. Why your new cancer biomarker may never work: recurrent patterns and remarkable diversity in biomarker failures. *Cancer Res* 2012;72:6097–6101.
- 11 Director's Challenge Consortium for the Molecular Classification of Lung A, Shedden K, Taylor JM, *et al*. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nat Med* 2008;14:822–827.
- 12 Tang H, Xiao G, Behrens C, *et al*. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *Clin Cancer Res* 2013;19:1577–1586.
- 13 Zhu CQ, Ding K, Strumpf D, *et al*. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. *J Clin Oncol* 2010;28:4417–4424.
- 14 Gold KA, Kim ES, Liu DD, *et al*. Prediction of survival in resected non-small cell lung cancer using a protein expression-based risk model: implications for personalized chemoprevention and therapy. *Clin Cancer Res* 2014;20:1946–1954.
- 15 D'Amico TA, Massey M, Herndon JE 2nd, *et al*. A biologic risk model for stage I lung cancer: immunohistochemical analysis of 408 patients with the use of ten molecular markers. *J Thorac Cardiovasc Surg* 1999;117:736–743.
- 16 Micke P, Mattsson JS, Djureinovic D, *et al*. The impact of the fourth edition of the WHO classification of lung tumours on histological classification of resected pulmonary NSCCs. *J Thorac Oncol* 2016;11: 862–872.
- 17 Tran L, Mattsson JS, Nodin B, *et al*. Various antibody clones of napsin A, thyroid transcription factor 1, and p40 and comparisons with cytokeratin 5 and p63 in histopathologic diagnostics of non-small cell lung

- carcinoma. *Appl Immunohistochem Mol Morphol* 2015;24:648–659.
- 18 Djureinovic D, Hallström BM, Horie M, *et al*. Profiling cancer testis antigens in non-small-cell lung cancer. *JCI Insight* 2016;1:e86837.
 - 19 Goldstraw P, Crowley J, Chansky K, *et al*. The IASLC Lung Cancer Staging Project: proposals for the revision of the TNM stage groupings in the forthcoming (seventh) edition of the TNM Classification of malignant tumours. *J Thorac Oncol* 2007;2:706–714.
 - 20 Travis WD, Brambilla E, Muller-Hermelink HK, *et al*. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart. IARC Press: Lyon, France, 2004: Chapter 1, pp 25–51.
 - 21 Goto A, Niki T, Chi-Pin L, *et al*. Loss of TSLC1 expression in lung adenocarcinoma: relationships with histological subtypes, sex and prognostic significance. *Cancer Sci* 2005;96:480–486.
 - 22 Hou J, Aerts J, den Hamer B, *et al*. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE* 2010;5:e10312.
 - 23 Xie Y, Xiao G, Coombes KR, *et al*. Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clin Cancer Res* 2011;17:5705–5714.
 - 24 Rousseaux S, Debernardi A, Jacquiau B, *et al*. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci Transl Med* 2013;5:186ra66.
 - 25 Okayama H, Kohno T, Ishii Y, *et al*. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res* 2012;72:100–111.
 - 26 Bild AH, Yao G, Chang JT, *et al*. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature* 2006;439:353–357.
 - 27 Raponi M, Zhang Y, Yu J, *et al*. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Res* 2006;66:7466–7472.
 - 28 Der SD, Sykes J, Pintilie M, *et al*. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. *J Thorac Oncol* 2014;9:59–64.
 - 29 Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B* 1995;57:289–300.
 - 30 Edlund K, Lindskog C, Saito A, *et al*. CD99 is a novel prognostic stromal marker in non-small cell lung cancer. *Int J Cancer* 2012;131:2264–2273.
 - 31 Therneau T. A Package for Survival Analysis in S. Available from <http://CRAN.R-project.org/package=survival> (accessed on 2015).
 - 32 Harrell FE Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361–387.
 - 33 Scholzen T, Gerdes J. The Ki-67 protein: from the known and the unknown. *J Cell Physiol* 2000;182:311–322.
 - 34 Kloppel G, Couvelard A, Perren A, *et al*. ENETS Consensus Guidelines for the Standards of Care in Neuroendocrine Tumors: towards a standardized approach to the diagnosis of gastroenteropancreatic neuroendocrine tumors and their prognostic stratification. *Neuroendocrinology* 2009;90:162–166.
 - 35 Yerushalmi R, Woods R, Ravdin PM, *et al*. Ki67 in breast cancer: prognostic and predictive potential. *Lancet Oncol* 2010;11:174–183.
 - 36 Jakobsen JN, Sorensen JB. Clinical impact of ki-67 labeling index in non-small cell lung cancer. *Lung Cancer* 2013;79:1–7.
 - 37 Nakamura N, Miyagi E, Murata S, *et al*. Expression of thyroid transcription factor-1 in normal and neoplastic lung tissues. *Mod Pathol* 2002;15:1058–1067.
 - 38 Sheppard MN. Specific markers for pulmonary tumours. *Histopathology* 2000;36:273–276.
 - 39 Barlesi F, Pinot D, Legoffic A, *et al*. Positive thyroid transcription factor 1 staining strongly correlates with survival of patients with adenocarcinoma of the lung. *Br J Cancer* 2005;93:450–452.
 - 40 Barletta JA, Perner S, Iafrate AJ, *et al*. Clinical significance of TTF-1 protein expression and TTF-1 gene amplification in lung adenocarcinoma. *J Cell Mol Med* 2009;13:1977–1986.
 - 41 Ren G, Baritaki S, Marathe H, *et al*. Polycomb protein EZH2 regulates tumor invasion via the transcriptional repression of the metastasis suppressor RKIP in breast and prostate cancer. *Cancer Res* 2012;72:3091–3104.
 - 42 Jiang T, Wang Y, Zhou F, *et al*. Prognostic value of high EZH2 expression in patients with different types of cancer: a systematic review with meta-analysis. *Oncotarget* 2016;7:4584–4597.
 - 43 Behrens C, Solis LM, Lin H, *et al*. EZH2 protein expression associates with the early pathogenesis, tumor progression, and prognosis of non-small cell lung carcinoma. *Clin Cancer Res* 2013;19:6556–6565.
 - 44 Kunju LP, Cookingham C, Toy KA, *et al*. EZH2 and ALDH-1 mark breast epithelium at risk for breast cancer development. *Mod Pathol* 2011;24:786–793.
 - 45 Kuramochi M, Fukuhara H, Nobukuni T, *et al*. TSLC1 is a tumor-suppressor gene in human non-small-cell lung cancer. *Nat Genet* 2001;27:427–430.
 - 46 Zeng D, Wu X, Zheng J, *et al*. Loss of CADM1/TSLC1 expression is associated with poor clinical outcome in patients with esophageal squamous cell carcinoma. *Gastroenterol Res Pract* 2016;2016:6947623.
 - 47 Zhang J, Ning J, Geng J, *et al*. Down-regulation of tumor suppressor in lung cancer 1 (TSLC1) expression correlates with poor prognosis in patients with colon cancer. *J Mol Histol* 2012;43:715–721.
 - 48 Ooi AT, Gomperts BN. Molecular pathways: targeting cellular energy metabolism in cancer via inhibition of SLC2A1 and LDHA. *Clin Cancer Res* 2015;21:2440–2444.
 - 49 Barron CC, Bilan PJ, Tsakiridis T, *et al*. Facilitative glucose transporters: Implications for cancer detection, prognosis and treatment. *Metabolism* 2016;65:124–139.
 - 50 Shiba M, Kohno H, Kakizawa K, *et al*. Ki-67 immunostaining and other prognostic factors including tobacco smoking in patients with resected nonsmall cell lung carcinoma. *Cancer* 2000;89:1457–1465.
 - 51 Younes M, Brown RW, Stephenson M, *et al*. Overexpression of Glut1 and Glut3 in stage I nonsmall cell lung carcinoma is associated with poor survival. *Cancer* 1997;80:1046–1051.
 - 52 Demarchi LM, Reis MM, Palomino SA, *et al*. Prognostic values of stromal proportion and PCNA, Ki-67, and p53 proteins in patients with resected adenocarcinoma of the lung. *Mod Pathol* 2000;13:511–520.
 - 53 Ko E, Lee BB, Kim Y, *et al*. Association of RASSF1A and p63 with poor recurrence-free survival in

- node-negative stage I-II non-small cell lung cancer. *Clin Cancer Res* 2013;19:1204–1212.
- 54 Masuda K, Takano A, Oshita H, *et al*. Chondrolectin is a novel diagnostic biomarker and a therapeutic target for lung cancer. *Clin Cancer Res* 2011;17:7712–7722.
- 55 Beer DG, Kardia SL, Huang CC, *et al*. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nat Med* 2002;8:816–824.
- 56 Hayes DN, Monti S, Parmigiani G, *et al*. Gene expression profiling reveals reproducible human lung adenocarcinoma subtypes in multiple independent patient cohorts. *J Clin Oncol* 2006;24:5079–5090.
- 57 Wistuba II, Behrens C, Lombardi F, *et al*. Validation of a proliferation-based expression signature as prognostic marker in early stage lung adenocarcinoma. *Clin Cancer Res* 2013;19:6261–6271.
- 58 Kratz JR, He J, Van Den Eeden SK, *et al*. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. *Lancet* 2012;379:823–832.
- 59 Nakagawa T, Toyazaki T, Chiba N, *et al*. Prognostic value of body mass index and change in body weight in postoperative outcomes of lung cancer surgery. *Interact Cardiovasc Thorac Surg* 2016;23:560–566.
- 60 Yotsukura M, Ohtsuka T, Kaseda K, *et al*. Value of the glasgow prognostic score as a prognostic factor in resectable non-small cell lung cancer. *J Thorac Oncol* 2016;11:1311–1318.
- 61 Ladanyi M, Pao W. Lung adenocarcinoma: guiding EGFR-targeted therapy and beyond. *Mod Pathol* 2008;21(Suppl 2):S16–S22.
- 62 Uchino K, Ito A, Wakayama T, *et al*. Clinical implication and prognostic significance of the tumor suppressor TSLC1 gene detected in adenocarcinoma of the lung. *Cancer* 2003;98:1002–1007.

Supplementary Information accompanies the paper on Modern Pathology website (<http://www.nature.com/modpathol>)