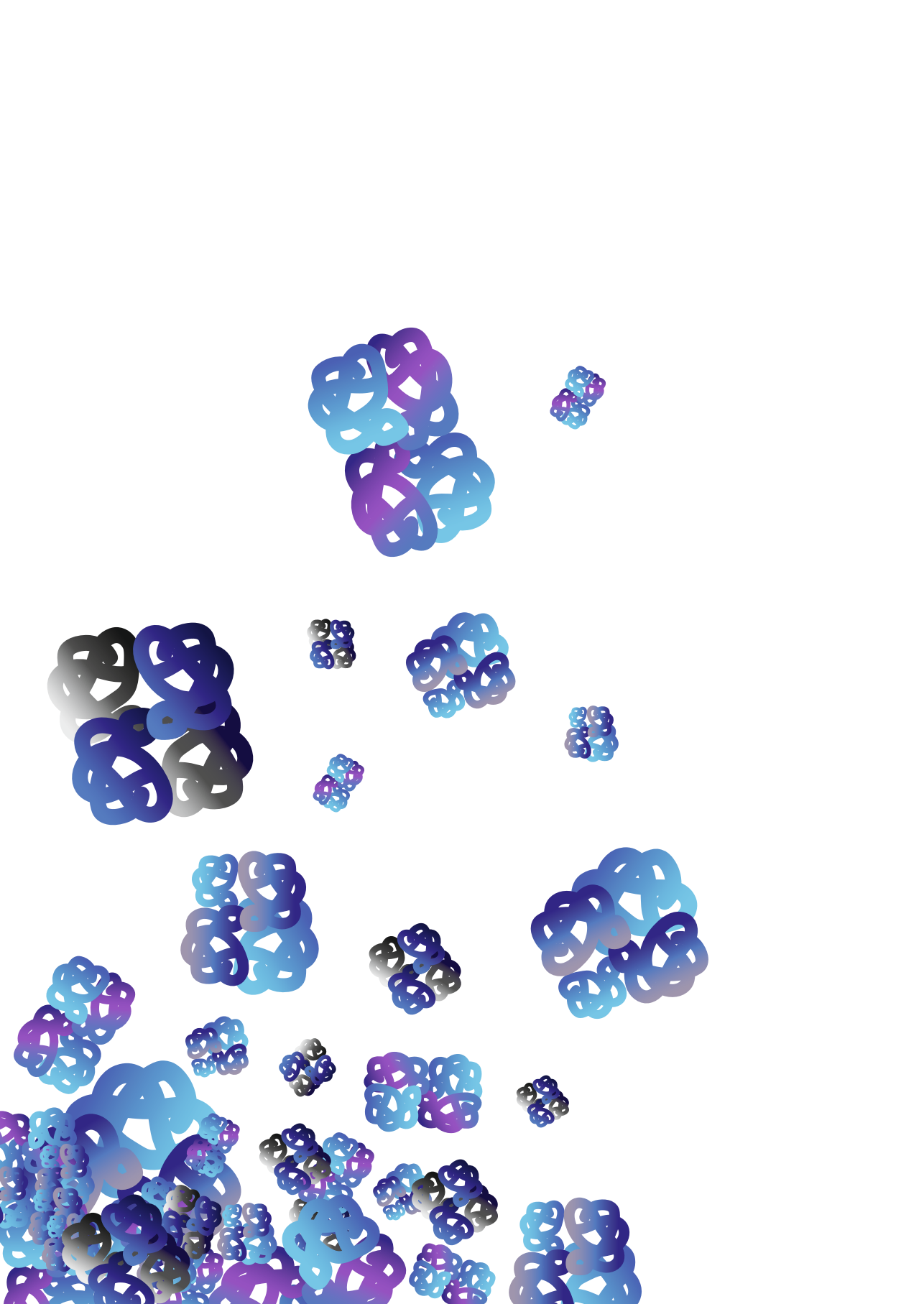
Biomarkers for the response to immunotherapy in patients with non-small cell lung cancer

Muller, M.



4 A serum protein classifier identifying patients with advanced non-small cell lung cancer who derive clinical benefit from treatment with immune checkpoint inhibitors

M. Muller*, K. Hummelink*, D.P. Hurkmans, A.N Niemeijer, K. Monkhorst, J. Roder, C. Oliveira, H. Roder, J.G. Aerts, E.F. Smit

Clin Cancer Res October 1 2020 (26) (19) 5188-5197

Abstract

Purpose: Pretreatment selection of patients with non–small cell lung cancer (NSCLC) who would derive clinical benefit from treatment with immune checkpoint inhibitors (CPIs) would fulfill an unmet clinical need by reducing unnecessary toxicities from treatment and result in substantial health care savings.

Experimental Design: In a retrospective study, mass spectrometry (MS)-based proteomic analysis was performed on pretreatment sera derived from patients with advanced NSCLC treated with nivolumab as part of routine clinical care (*n* = 289). Machine learning combined spectral and clinical data to stratify patients into three groups with good ("sensitive"), intermediate, and poor ("resistant") outcomes following treatment in the second-line setting. The test was applied to three independent patient cohorts and its biology was investigated using protein set enrichment analyses (PSEA).

Results: A signature consisting of 274 MS features derived from a development set of 116 patients was associated with progression-free survival (PFS) and overall survival (OS) across two validation cohorts (*N* = 98 and *N* = 75). In pooled analysis, significantly better OS was demonstrated for "sensitive" relative to "not sensitive" patients treated with nivolumab; HR, 0.58 (95% confidence interval, 0.38–0–87; *P* = 0.009). There was no significant association with clinical factors including PD-L1 expression, available from 133 of 289 patients. The test demonstrated no significant association with PFS or OS in a historical cohort (*n* = 68) of second-line NSCLC patients treated with docetaxel. PSEA revealed proteomic classification to be significantly associated with complement and wound-healing cascades.

Conclusions: This serum-derived protein signature successfully stratified outcomes in cohorts of patients with advanced NSCLC treated with second-line PD-1 CPIs and deserves further prospective study.

Translational Relevance

Predictive biomarkers for the efficacy of PD-L1 inhibition in non–small cell lung cancer (NSCLC) beyond PD-L1 are lacking. We retrospectively developed a pretreatment proteomic signature derived from peripheral blood that was able to stratify patients for benefit of nivolumab in treatment of relapsed NSCLC. A signature consisting of 274 mass spectral features derived from a development set of 116 patients was associated with progression-free survival and overall survival (OS) across two validation cohorts (*N* = 98 and *N* = 75). In pooled analysis, a significantly better OS was demonstrated for "sensitive" relative to "not sensitive" patients, HR 0.58 (95% confidence interval, 0.38–0.87; *P* = 0.009). There was no significant association with clinical factors including PD-L1 IHC. Further prospective exploration of the predictive capabilities of this assay is underway.

1. Introduction

The addition of immune checkpoint inhibitors (CPIs) to the armamentarium of medical treatment of advanced non–small cell lung cancer (NSCLC) has increased survival for a minority of patients. Historically, in patients with metastatic disease, 2-year survival rates following platinum-based chemotherapy were 10%–20% [198]. In recent phase III studies, either comparing CPIs alone or CPI chemotherapy to chemotherapy[199], 2-year survival rates in the CPI arms range from 32% to 67%. In addition, long-term follow-up of patients treated in early single-agent CPI studies indicates that 5-year survival of 15%–20% may be expected, even in heavily pretreated patients [40, 200].

At the same time, it is clear that not all patients benefit from treatment with CPIs. Indeed, response rates and survival times can be augmented by pretreatment selection based on tumor characteristics such as PD-L1 expression [40], staining of CD8-positive cells[201], tumor mutational burden (TMB) [202], and other genomic markers [203, 204]. The predictive power of the best studied of these PD-L1 IHC is far from perfect. For example, in patients with previously treated NSCLC with PD-L1 staining of at least 50%, the objective response rate to pembrolizumab was 44% [40]. Thus, alternative predictive biomarkers for response and clinical benefit are needed. We sought to develop a serum-based, pretreatment protein test to avoid the need for tissue biopsies, which are typically required to analyze tumor-related biomarkers. Here, we report on the development of such a test in advanced NSCLC treated with single-agent CPI in the second-line setting.

2. Materials and Methods

2.1 Patient cohorts and sample sets

Pretreatment serum samples, collected within 1 month of immunotherapy initiation, were available from four cohorts of patients. The development set consisted of 116 samples from patients treated at the Netherlands Cancer Institute (Amsterdam, the Netherlands) between May 2015 and March 2017. Validation set 1 consisted of 98 samples from patients treated at the Vrije Universiteit Medical Center (Amsterdam, the Netherlands) or the Netherlands Cancer Institute (Amsterdam, the Netherlands) between June 2015 and July 2018. Validation set 2 comprised samples from 75 patients treated at the Erasmus University Medical Centre (Rotterdam, the Netherlands) between April 2016 and July 2017. Patients, identified according to criteria established in the phase III trials demonstrating benefit for nivolumab over docetaxel [3, 5], received nivolumab 3 mg/kg, administered as an intravenous infusion, every 2 weeks, for advanced NSCLC after platinum-containing chemotherapy as part of routine clinical care. Patients in the development cohort and validation set 2 were treated in second-line. Validation set 1 contained 58 patients treated in second-line and 40 patients treated in higher lines. The cohorts comprised all patients in the respective institutions who provided pretreatment serum samples available for analysis, were eligible for immunotherapy as routine care, and who received at least one dose of nivolumab. Response to treatment was evaluated according to RECIST v1.1 every 6 weeks for the first 12 weeks and every 3 months thereafter. In addition, a fourth cohort of pretreatment serum samples (chemotherapy cohort) was collected from patients with advanced NSCLC treated in second-line with chemotherapy while enrolled in a clinical trial (NCT00989690)[205].

Samples were available for 68 of the 74 patients who received docetaxel (75 mg/m2 every 21 days) in this study. Trial inclusion and exclusion criteria have been published elsewhere [205]. All samples were obtained in the context of biobanking protocols or a clinical trial for which institutional review board approval was sought and obtained. All patients provided written informed consent according to local ethical standards and adhered to standards set out in the Declaration of Helsinki. Progression-free survival (PFS) was measured from start of treatment until progression of disease, death, or loss to follow-up. Overall survival (OS) was defined as time from start of therapy until death or loss to follow-up.

2.2 PD-L1 IHC

Tumor PD-L1 expression scoring was performed according to the instruction manual of the qualitative IHC assay developed as a complementary diagnostic tool for nivolumab (PD-L1 IHC 22C3 pharmDx, Dako). PD-L1 expression levels were determined by observing complete circumferential or partial linear expression (at any intensity) of PD-L1 on the plasma cell membrane of viable tumor cells. In parallel, the pattern of staining in CD4-stained slides, which also stain CD4+ lymphocytes and macrophages, was evaluated and compared with PD-L1–stained slides to avoid false positive assessment due to PD-L1–expressing macrophages in between tumor cells. Assessment of expression levels was performed in sections that included at least 100 tumor cells that could be evaluated.

2.3 Spectral acquisition and processing

Samples were processed using standardized operating procedures. We used the Deep matrix-assisted laser desorption/ionization (MALDI) method of mass spectrometry on a MALDI Time-of-Flight Mass Spectrometer (SimulTof Systems) to generate reproducible mass spectra from small amounts of serum (3 µL) [206]. This approach reveals mass spectral (MS) peaks with a greater dynamic range than previously possible by exposing the samples to 400,000 MALDI laser “shots,” rather than the several thousand used in standard applications. The spectra were processed to render them comparable between patients, and 274 MS features (peaks) were selected for further analysis for their known reproducibility and stability (listed in Supplementary Data). Sample processing and MS analysis followed methods presented previously [207, 208] and are outlined in the Supplementary Materials and Methods. Parameters for these procedures were established using only the 116-sample development set, and this fixed procedure was applied to all other sample sets without modification.

2.4 Test development

Test development was carried out using the Diagnostic Cortex platform[209], which has been used previously to design tests that were able to stratify patients by outcome in various settings, for example, to identify patients with advanced melanoma likely to be sensitive to CPIs [207, 208]. The approach incorporates machine learning concepts and elements of deep learning[210] to facilitate test development in cases where there are more measured attributes than samples. The potential for overfitting was minimized, thus allowing the creation of tests that can generalize to unseen datasets. Tests were created averaging over many splits of the development set into training and test sets,

and reliable test performance estimates can be obtained from the development set by restricting averages to the test set evaluations ("out-of-bag estimates") [211].

For successful supervised learning, suitable training class labels are required. We used a semisupervised approach [212] that does not require accurate prespecification of patients into better or worse outcome training classes and allows us to be guided by the gold standard time-to-event outcomes of OS and PFS. An approximation was made for training classes, with patients with the lowest time-to-event outcome times assigned to the "negative" class and those with the highest time-to-event outcome times assigned to the "positive" class. A classifier was constructed using these training classes and used to generate classifications for all samples in the development set using out-of-bag evaluations. These resulting classifications were then used as improved training class labels for a second iteration of classifier construction. This simultaneous iterative refinement of the classifier and the classes used in training generally converges quickly and reveals the underlying structure of the MS data and its association with clinical outcomes [212]. Full details of the application of the method in this setting are provided in the Supplementary Materials and Methods.

One classifier previously developed with the Diagnostic Cortex platform was used as part of the developed test. BDX008 was created to stratify patients with advanced melanoma into groups with better and worse outcomes when treated with nivolumab [208]. It has been validated in multiple independent cohorts of patients with melanoma treated with CPIs [208, 213]. Also, it has demonstrated some ability to stratify OS of patients with advanced NSCLC treated with nivolumab [214]. A version of BDX008, adapted for the spectral preprocessing parameters and feature definitions in this project, was created (see Supplementary Data: Supplementary Materials and Methods for details).

Preliminary statistical considerations showed a binary split of the development set into equal-sized groups would allow detection of a HR between the groups of 0.55 with 90% power, assuming fully mature clinical data and a significance level of 95%. Similar considerations for a ternary split into equal size subgroups would allow detection of an HR of 0.48 under the same specifications.

All reference data and test parameters were generated solely using the development set. Validation sets and the chemotherapy cohort were never used in the creation of any components of the test. All elements of the classification algorithms were locked prior to running the test on the validation sets and chemotherapy cohort.

2.5 Protein set enrichment analysis

This analysis applies the gene set enrichment analysis (GSEA) method [215] to protein expression data. The method identifies expression differences that are consistent across prespecified groups or sets of attributes, in this case, sets of proteins that are associated with particular biological processes. Two additional independent reference sets of serum samples with matched MS data and protein expression data were used for this set enrichment analysis. One sample set was composed of 49 samples with protein expression data from a panel of 1,129 proteins; the second set had 100 samples

with protein expression data from a panel of 1,305 proteins (protein expression measurements were generated by SomaLogic). Specific protein sets were created as the intersection of the list of the panel targets and results of queries for biological functions from gene ontology, using AmiGO2 tools (http://amigo.geneontology.org/amigo) and UniProt databases (https://www.uniprot.org/). The protein set enrichment analyses (PSEA) method associated test classification with these biological functions via a rank-based correlation of the measured protein expressions with the test classifications of the reference samples [216]. The mass spectral features associated with biological processes (in particular immune response type 2) were determined using Spearman correlation of the measured protein expressions with the mass spectral features [216] using the 49-sample reference set only. While the implementation closely follows the GSEA approach, we employed an extension of the standard method that increases the statistical power to detect associations between phenotype (test classification subgroup) and biological process[217]. The PSEA was carried out using a C# implementation and MATLAB (MathWorks). PSEA P values were defined as described by Subramanian and colleagues[213]. FDRs for the PSEA calculations were assessed using the method of Benjamini and Hochberg [218].

2.6 Other statistical analysis

All analyses, except the PSEA, were carried out using SAS9.3 (SAS Institute) or PRISM (GraphPad). Survival/PFS plots and medians were generated using the Kaplan–Meier method. Association between test classification and categorical or continuous variables was assessed using Fisher exact test and Mann–Whitney test, respectively. All P values are two-sided.

Table 1 - Patient characteristics and outcomes for all cohorts: development, validation set 1, validation set 2, and chemotherapy cohort.

		Development (N=116)	Validation 1 (N=98)	Validation 2 (N=75)	Chemotherapy (N=68)
Age Median (Range)		**65 (43-83)**	**64 (29-77)**	**65 (35-78)**	**64 (39-77)**
Gender, N (%)	Male	66 (57)	51 (52)	48 (64)	52 (76)
	Female	50 (43)	47 (48)	27(36)	16 (24)
PS, n (%)	0	36 (32)	20 (20)	18 (32)	35 (51)
	1	60 (54)	65 (66)	37 (66)	29 (43)
	2+	15 (14)	13 (13)	1 (2)	4 (6)
Smoking Status, n (%)	Ever	104 (91)	88 (92)	61 (92)	64 (94)
	Never	10 (9)	8 (8)	5 (8)	4 (6)
Histology, n (%)	Adenocarcinoma	77 (66)	42 (74)	49 (65)	47 (75)
	Squamous	26 (22)	10 (18)	17 (23)	12 (19)
	Other	13 (11)	5 (9)	9 (12)	4 (6)
PD-L1 expression, n (%)	≥1%	33 (28)	12 (14)	16(21)	0 (0)
	<1%	43 (37)	30 (29)	9 (12)	0 (0)
	NA	40 (34)	56 (57)	50 (67)	68 (100)

Table 1 - Patient characteristics and outcomes for all cohorts: development, validation set 1, validation set 2, and chemotherapy cohort. *(Continued)*

		Development (N=116)	Validation 1 (N=98)	Validation 2 (N=75)	Chemotherapy (N=68)
Response, n (%)	CR	1 (1)	1 (0)	0 (0)	0 (0)
	PR	16 (14)	28 (28)	15 (20)	7 (10)
	SD	19 (16)	26 (33)	25 (33)	23 (34)
	PD	65 (56)	37 (33)	31 (41)	22 (32)
	NA/NE	15 (13)	6 (7)	4 (5)	16 (24)
PFS (months)	Median	2.6	4.1	4.3	3.5
OS (months)	Median	8.5	8.4	12.0	8.0

3. Results

Patient characteristics and overall outcomes for all four cohorts are summarized in Table 1 and were typical of patients with advanced NSCLC treated predominantly in the second-line setting. Clinicopathologic characteristics were generally similar between the four cohorts, although the proportion of patients with performance status 2 or higher was larger in the development cohort and validation set 1, and the proportion of patients with performance status 0 was higher in the chemotherapy cohort. PD-L1 status was not available for the chemotherapy cohort and was missing for at least one-third of patients in the other three cohorts.

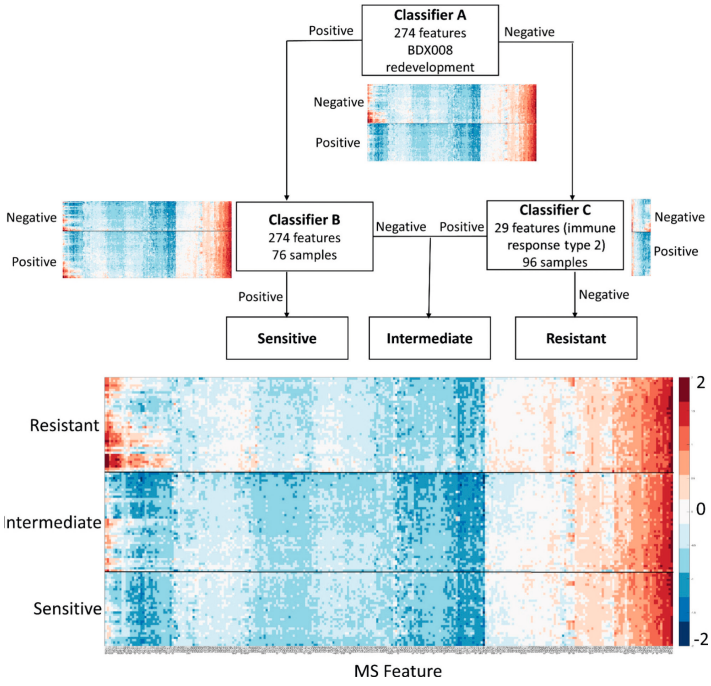


Figure 1: Schema showing how the final test result is produced from the three classifiers A, B, and C.

Notes: *Heatmaps within the schema show* log_{10} values of features used in each classifier (x-*axis) for the development cohort of 116 samples, grouped by individual classifier results, negative or positive. The heatmap below the schema shows the* log_{10} values of all 274 features used within the test for all samples in the development cohort, grouped by test classification (resistant, intermediate, or sensitive).

3.1 Development of the test

A ternary test was developed that was able to stratify the development set of 116 samples into three groups with different outcomes after anti-PD-1 treatment, that is, the resistant group (with poor outcomes), the intermediate group (with intermediate outcomes), and the sensitive group (with good outcomes). The ternary test result was generated by combining the results of three binary classification algorithms (classifiers). Each of the three classifiers stratified patients into two groups: "positive" with better outcomes and "negative" with worse outcomes. The binary results were integrated as shown in Figure 1 to yield the final test result. First, classifications were generated for all samples by classifier A, the version of the preexisting BDX008 test adapted to the spectral processing used in this project. To identify a group of patients least likely to have good outcomes, the patients classified as negative by classifier A were subsequently classified by classifier C. This classifier was developed using the subset of MS features found to be associated with immune response type 2 by PSEA and a subset of the development cohort enriched for inferior outcomes, by excluding patients designated as BDX008+ and having performance status 0 (the MS features in this subset are listed in the Supplementary Materials and Methods). Samples designated as negative by both

classifier A and classifier C were classified as "resistant." To identify a group of patients likely to have the best outcomes, the patients classified as positive by classifier A were further classified by classifier B. This classifier was developed using all 274 mass spectral features on a subset of the development set enriched for better outcomes, by excluding patients who were classified both as BDX008− and negative by classifier C. Samples designated positive by both classifier A and classifier B were classified as "sensitive." All samples not classified as "sensitive" or "resistant" were classified as "intermediate." More details of the test development process and parameters are provided in the Supplementary Data. Reproducibility was assessed by running the test on the 98 serum samples of validation set 1 twice, 13 months apart. Concordance between classifications was 85%. For identification of patients with resistant outcomes (resistant vs. not resistant, i.e., sensitive and intermediate), concordance was 91% and for identification of patients with sensitive outcomes (sensitive vs. not sensitive, i.e., resistant and intermediate), concordance was 93%.

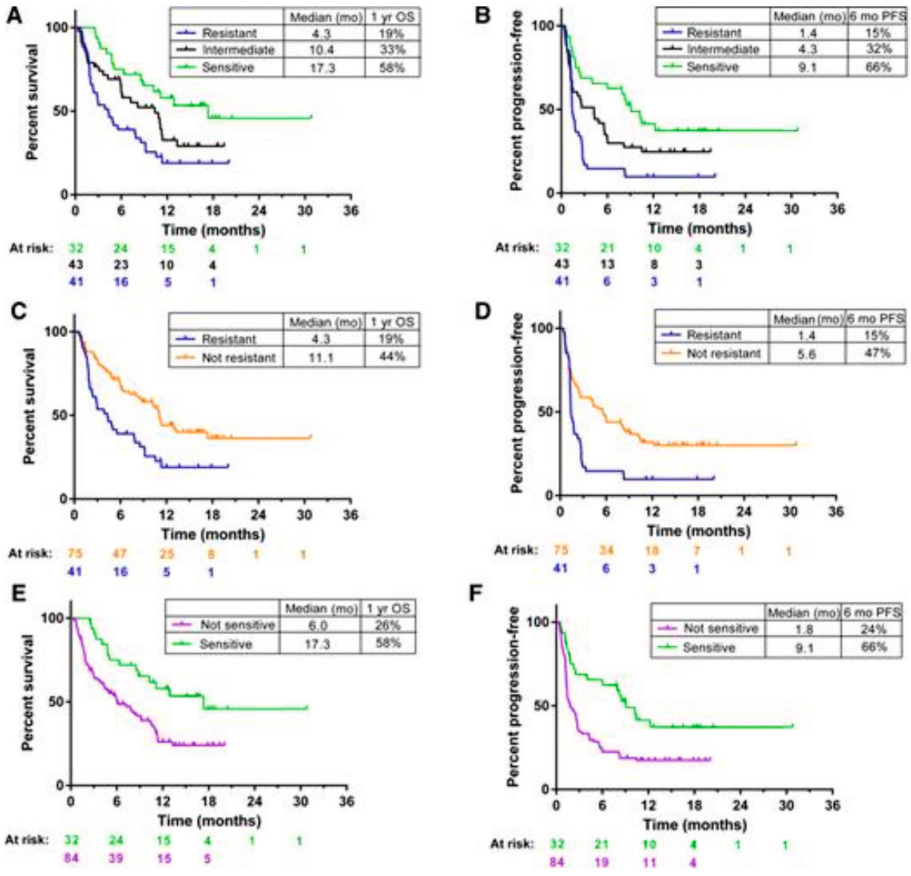


Figure 2 - Outcomes by test classification for the development cohort.

The test was able to stratify patients into three groups (sensitive, intermediate, and resistant) with different OS and PFS. Of the 116 samples in the development set, 41 (35%) were classified as resistant, 43 (37%) as intermediate, and 32 (28%) as sensitive. Kaplan–Meier plots of OS and PFS by classification groups are shown in Figure 2A and 2B. PFS for the resistant subgroup was significantly shorter than for the other groups [resistant vs. sensitive: HR, 0.33 (95% confidence interval (CI), 0.19–0.58); P < 0.001 and resistant vs. intermediate: HR, 0.59 (95% CI, 0.37–0.96); P = 0.035]. Median PFS was 1.4 (95% CI, 1.3–2.3) months for the resistant group, 4.3 (95% CI, 1.4–5.7) months for the intermediate group, and 9.1 (95% CI, 2.5–undefined) months for the sensitive group. OS for the resistant subgroup was significantly shorter than for the sensitive subgroup and numerically shorter than for the intermediate group (resistant vs. sensitive: HR, 0.34 (95% CI, 0.19–0.64); P < 0.001 and resistant vs. intermediate: HR, 0.63 (95% CI, 0.38–1.06); P = 0.083]. Median OS was 4.3 (95% CI, 2.0–7.9) months for the resistant subgroup, 10.4 (95% CI, 5.9–11.4) months for the intermediate group, and 17.3 (95% CI, 8.5–undefined) months for the sensitive group. Test classification was also associated with response (P < 0.001, see Supplementary Data: Supplementary Results; Supplementary Table S12). Eighty-five percent of patients classified as resistant experienced progressive disease as best response and only 10% had a response (all partial). In the sensitive group, only 28% of patients had progressive disease as best response and 28% achieved a response (one complete response (CR) and eight partial responses (PR) as best response of 32 patients).

For differentiating patients with the worst outcome from the remainder of the cohort, we compared the resistant subgroup with the "not resistant" group, that is, the combination of intermediate and sensitive subgroups, see Figure 2C and D. The resistant subgroup had significantly inferior OS and PFS than the other subgroups [HR, 0.48 (95% CI, 0.30–0.77); P = 0.002 for OS and HR, 0.46 (95% CI, 0.30–0.71); P < 0.001 for PFS]. These differences remained significant for PFS (P = 0.015) and trended to significance for OS (P = 0.062) in multivariate analysis when adjusted for other baseline characteristics, including performance status and PD-L1 expression.

The patients with the best outcomes (sensitive subgroup) were compared with the "not sensitive" group, that is, the remainder of the cohort (resistant + intermediate subgroups; Figure 2E and F). Patients classified as sensitive had significantly better OS and PFS than patients classified as not sensitive [HR, 0.45 (95% CI, 0.25–0.79); P = 0.006 for OS and HR, 0.45 (95% CI, 0.27–0.76); P = 0.003 for PFS]. Median OS was 17.3 (95% CI, 8.5–undefined) months for the sensitive group, compared with 6.0 (95% CI, 4.3–9.2) months for the not sensitive group; median PFS was 9.1 (95% CI, 2.5–undefined) months for the sensitive group, compared with only 1.8 (95% CI, 1.4–2.7) months for the not sensitive group. In multivariate analyses, while the effect sizes for OS and PFS remained substantial (HR, 0.60 and 0.63, respectively), classification of sensitive versus not sensitive did not retain its independent significance as a predictive factor (Supplementary Data: Supplementary Results; Supplementary Tables S13 and S14).

Baseline patient characteristics showed no association with test classification for P < 0.05 (Supplementary Data: Supplementary Results; Supplementary Table S15). In particular,

PD-L1 expression was not significantly correlated with test classification (P = 0.387 for ternary classification vs. PD-L1+/PD-L1−/NA).

3.2 Validation

The locked test was applied to samples from validation sets 1 and 2 and the chemotherapy cohort. Validation set 1 had been used in a previous investigation [219] and, therefore, while it was not used in test development, validation set 1 could not be run blinded to clinical data. The chemotherapy cohort was a subset of a previously analyzed clinical trial comparing chemotherapy and targeted therapy, and hence, could also not be tested blinded to clinical data. Testing of validation set 2 was completely blinded to all clinical data. Statistical consideration of power to detect the effect sizes observed in the development cohort for each validation set and the chemotherapy cohort is outlined in the Supplementary Data.

Within the validation sets, the number and proportions of patients assigned to each classification group were 37 (38%)/32 (43%) resistant, 30 (31%)/19 (25%) intermediate, and 31 (32%)/24 (32%) sensitive for set 1/set 2, respectively. Kaplan–Meier plots of OS by test classification, resistant versus not resistant and sensitive versus not sensitive, are shown for the validation sets in Figure 3A–3D. In validation set 1, Figure 3A and 3B, patients classified as resistant had significantly worse OS than not resistant patients [HR, 0.60 (95% CI, 0.37–0.97); P = 0.037] and patients classified as sensitive had significantly better OS than not sensitive patients [HR, 0.56 (95% CI, 0.33–0.97); P = 0.038]. One-year survival for the sensitive group was 65% and the corresponding median was 15.3 (95% CI, 8.8–undefined) months. In contrast, median OS was only 4.8 (95% CI, 2.9–9.3) months in the resistant group, with 29% OS at 1 year. PFS was numerically superior in the sensitive group and inferior in the resistant group, but the differences in outcome were smaller and did not reach statistical significance (see Supplementary Data: Supplementary Results; Supplementary Figs. S1 and S2). Analysis of the subgroup of patients treated with nivolumab in third- or higher-line (N = 40), showed similar behavior in OS and PFS, with resistant patients showing a trend to shorter outcomes [HR, 0.49 (95% CI, 0.23–1.04); P = 0.062 for OS and HR, 0.50 (95% CI, 0.25–1.02); P = 0.057 for PFS] and sensitive patients showing numerically longer survival [HR, 0.48 (95% CI, 0.21–1.10); P = 0.082 for OS and HR, 0.62 (95% CI, 0.31–1.23); P = 0.172 for PFS]. Kaplan–Meier plots for this subgroup are shown in the Supplementary Data.

Results for validation set 2 are shown in Figure 3C and 3D. Patients classified as resistant had worse OS than not resistant patients [HR, 0.39 (95% CI, 0.19–0.77); P = 0.007]. The comparison of OS between the sensitive group and the not sensitive patients yielded an HR of 0.58, but did not show a significant difference (P = 0.179). However, for ternary test classifications, the sensitive group had longer OS than the resistant group [HR, 0.41 (95% CI, 0.18–0.94); P = 0.036]. Full analysis for the sensitive/intermediate/resistant classifications can be found in Supplementary Data: Supplementary Results. Analysis of PFS showed only numerical differences between classification groups.

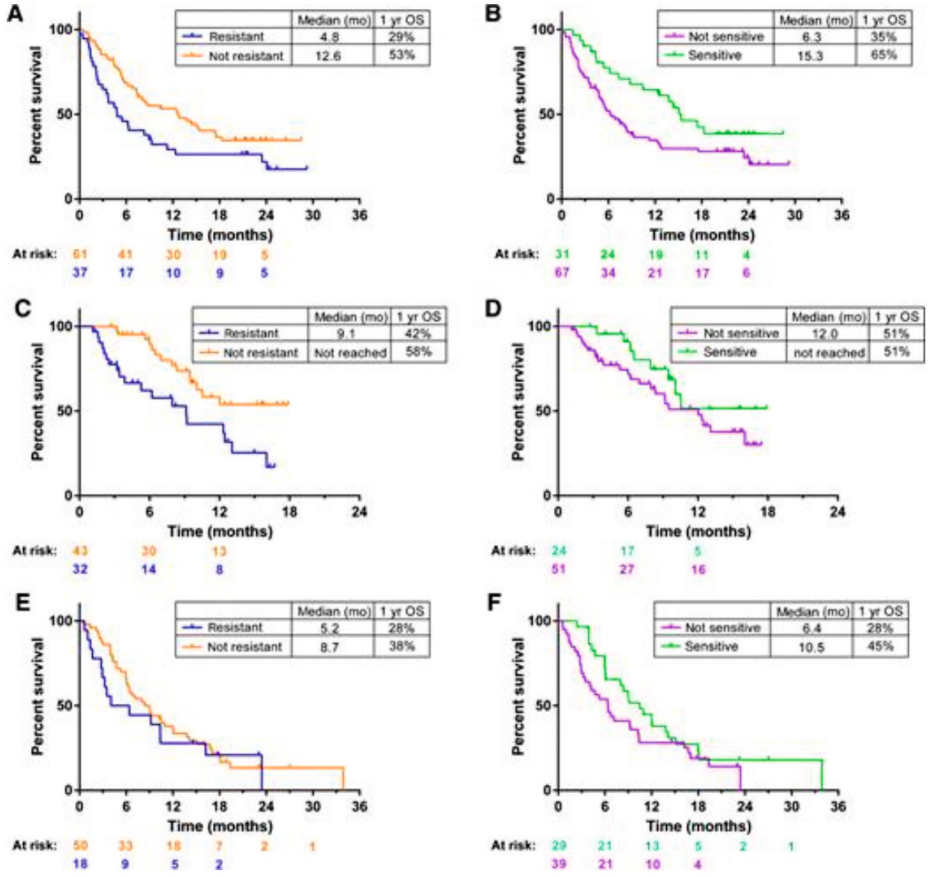


Figure 3 - Kaplan–Meier plots of OS for the validation sets and the chemotherapy cohort.
Notes: *A, Validation set 1 resistant versus not resistant. B, Validation set 1 sensitive versus not sensitive. C, Validation set 2 resistant versus not resistant. D, Validation set 2 sensitive versus not sensitive. E, Chemotherapy cohort sensitive versus not sensitive. F, Chemotherapy cohort resistant versus not resistant.*

As results were consistent across cohorts, within the limits of relatively small subgroup sizes, a pooled analysis of all patients treated in second-line setting with nivolumab was carried out stratified by cohort (N = 249). There was no indication of any correlation of PD-L1 expression with test classification (P = 0.292, 0.810, and 0.337 for ternary, resistant vs. not resistant, and sensitive vs. not sensitive test classifications), although positive PD-L1 expression was a predictor of improved OS and PFS in the pooled analysis [HR, 1.60 (95% CI, 1.01–2.54); P = 0.046 for OS and HR, 1.61 (95% CI, 1.07–2.44); P = 0.023 for PFS]. Indeed, analysis including test classification and PD-L1 expression demonstrated both to be independent predictors of PFS (see Supplementary Data). Within the pooled second-line population, multivariate analysis showed that the resistant versus not resistant stratification was a significant independent predictor of OS (P < 0.001) and PFS (P = 0.006) when adjusted for multiple baseline factors (Table 2). The sensitive versus not sensitive

stratification was a significant independent predictor of OS (P = 0.009) and showed a trend for prediction of PFS (P = 0.079).

Figure 3E and 3F show OS for classification groups obtained by applying the test to pretreatment samples of the chemotherapy cohort, in which patients received docetaxel as second-line therapy. There was no indication that the test was able to stratify patients by outcome following this single-agent chemotherapy (P = 0.471 and P = 0.165 for OS comparison of resistant vs. not resistant and sensitive vs. not sensitive, respectively).

Table 2 - Multivariate analysis of OS and PFS stratified by cohort for the pooled second-line population for test classification resistant versus not resistant (analysis 1) and test classification sensitive versus not sensitive (analysis 2).

		OS		PFS	
Analysis 1		**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
Test classification (vs resistant)	not resistant	0.52 (0.37-0.74)	<0.001	0.64 (0.47-0.88)	0.006
Histology (vs adeno)	squamous	0.83 (0.54-1.28)	0.403	1.10 (0.75-1.60)	0.639
	NA/other	1.10 (0.66-1.85)	0.711	1.09 (0.69-1.70)	0.718
Age (vs >=65)	<65	1.14 (0.80-1.63)	0.455	1.27 (0.93-1.73)	0.130
Gender (vs male)	female	0.52 (0.35-0.76)	0.001	0.69 (0.50-0.96)	0.027
PS (vs 0)	1	1.56 (1.02-2.39)	0.040	1.37 (0.96-1.97)	0.084
	2+	3.66 (2.00-6.67)	<0.001	2.30 (1.31-4.06)	0.004
	NA	2.29 (1.15-4.54)	0.018	1.90 (1.05-3.46)	0.035
Smoking (vs Ever)	never	1.87 (0.96-3.64)	0.064	1.47 (0.81-2.67)	0.209
	NA	0.76 (0.30-1.92)	0.559	0.76 (0.33-1.76)	0.521
PD-L1 (vs Positive)	negative	1.20 (0.74-1.94)	0.461	1.31 (0.85-2.03)	0.218
	NA	0.84 (0.52-1.36)	0.474	0.86 (0.57-1.30)	0.476
Analysis 2		**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
Test (vs not sensitive)	sensitive	0.58 (0.38-0.87)	0.009	0.73 (0.51-1.04)	0.079
Histology (vs adeno)	squamous	0.84 (0.54-1.30)	0.428	1.12 (0.76-1.63)	0.573
	NA/other	1.13 (0.68-1.87)	0.648	1.10 (0.70-1.71)	0.683
Age (vs >=65)	<65	1.06 (0.75-1.51)	0.750	1.21 (0.89-1.65)	0.227
Gender (vs male)	female	0.49 (0.33-0.72)	<0.001	0.66 (0.47-0.91)	0.011
PS (vs 0)	1	1.41 (0.92-2.17)	0.116	1.32 (0.92-1.90)	0.136
	2+	3.31 (1.78-6.13)	<0.001	2.19 (1.22-3.91)	0.008
	NA	2.25 (1.14-4.45)	0.020	1.95 (1.07-3.55)	0.028
Smoking (vs Ever)	never	1.82 (0.93-3.57)	0.082	1.48 (0.81-2.71)	0.205
	NA	0.83 (0.33-2.11)	0.693	0.84 (0.36-1.95)	0.676
PD-L1 (vs Positive)	negative	1.22 (0.75-1.98)	0.417	1.34 (0.87-2.07)	0.189
	NA	0.97 (0.60-1.55)	0.882	0.94 (0.62-1.41)	0.755

Table 3 - PSEA of test classifications resistant vs not resistant.

Biological Process	*p* value of association	FDR
Acute phase response	<0.0001	<0.002
Acute inflammatory response	0.0001	<0.002
Wound healing	0.0002	<0.002
Complement activation	0.0005	<0.003
Innate immune response	0.0014	<0.01
Chronic inflammatory response	0.0044	<0.02
Extra cellular matrix	0.0231	<0.08
IFN type 1	0.0315	<0.1
Cellular component morphogenesis	0.0317	<0.1
Immune tolerance and suppression	0.0526	<0.2
B cell mediated immunity	0.0526	<0.2
Angiogenesis	0.0753	<0.2
Natural Killer (NK) cell mediated immunity	0.1222	<0.3
Behavior	0.1270	<0.3
Cytokine production involved in immune response	0.3198	<0.5
Glycolysis and positive regulators	0.3560	<0.6
Epithelial-Mesenchymal Transition	0.4548	<0.6
Type17 immune response	0.4668	<0.6
Type1 immune response	0.5102	<0.7
Type2 immune response	0.7791	<0.9
Response to hypoxia	0.9287	<1
T cell mediated immunity	0.9861	<1
IFN-Gamma	0.9884	<1

3.3 Protein set enrichment

To examine the potential biological mechanisms underlying the test, the association of test classification with various biological processes was assessed using PSEA methods[215-217]. The results are summarized in Table 3. Acute phase response, acute inflammatory response, wound healing, and complement activation were identified as associated with test classification with $P < 0.001$. In addition, innate immune response and chronic inflammatory response were identified as associated with $P < 0.01$. Similar analysis was performed comparing the sensitive subgroup with the remaining patients. Only immune tolerance and suppression were identified as associated with test classification with $P < 0.01$ (FDR < 0.1). Full results for sensitive versus not sensitive phenotype are contained in the Supplementary Data: Supplementary Results; Supplementary Table S21.

4. Discussion

Here, we report the establishment of a pretreatment serum proteomic classifier that separates those patients who obtain little from those that obtain durable clinical benefit from treatment with the PD-1 inhibitor, nivolumab, as second-line treatment for advanced NSCLC. On the basis of 274 MS features, patients could be classified as being resistant, intermediate, or sensitive. The difference in OS between resistant and not resistant patients was highly significant: the HR was 0.48, and median survival times were 4.3 months versus 11.1 months, respectively. The test was validated while blinded to clinical outcome data with an independent set of patients with advanced NSCLC, treated at a different institution. The classifier failed to stratify outcomes within a historical cohort of patients with advanced NSCLC treated with docetaxel as second-line therapy. Moreover, test classification, as expected, was independent of well-established clinical factors and notably showed no evidence of association with PD-L1 expression.

A serum test would have obvious advantages, such as ease of detection using one blood draw. Also, the test may avoid the issue of intrapatient tumor heterogeneity and could assess host factors that are not captured by examination of the tumor microenvironment in histologic samples. Further characterization of the classifier revealed that the classification phenotypes identified are associated with biological processes known to confer a poor prognosis in lung cancer. Several lines of research indicate that complement, as a member of a diverse family of innate immune proteins, is involved in dysregulation of mitogenic signaling and escape from immune surveillance [220, 221]. Complement activation, as measured by Cd4, a stable complement degradation product, in serum of patients with early-stage NSCLC was significantly associated with poor prognosis [222]. A number of authors have identified the ratio of the acute phase protein, serum C-Reactive protein, to albumin as a negative prognostic factor in both early and advanced NSCLC[64]. Intratumoral wound-healing signatures, as measured by mRNA expression arrays, are considered to be T-cell suppressive and have been observed in several tumor types, among them NSCLC [63]. Interestingly, sera derived from patients with tumors exhibiting wound-healing signatures elicited identical signatures from nontumor-associated fibroblasts, which were found to be a powerful predictor of an unfavorable clinical course [223]. These observations may provide the biological basis of our findings, although a direct link between the abundance of these circulating proteins and absence of a response to PD-1 inhibitors remains to be established.

The results obtained in this study do not stand alone. Weber and colleagues identified a protein classifier from sera of patients with melanoma treated with PD-1 inhibitors, employing the same technology that was used in our study. This was validated in multiple patient cohorts treated with PD-1 inhibitors and CTLA4 antagonists [207]. As here, they were able to identify, prior to initiation of treatment, patients who had a favorable outcome following treatment. Biological processes associated with that classifier included complement, wound healing, and acute phase pathways, all upregulated in the poor prognosis group, corroborating our results. Further evidence that the pretreatment circulating proteome provides important information on checkpoint efficacy was provided in the context of a phase II study where atezolizumab was compared with

docetaxel as second-line treatment in 272 patients with advanced NSCLC [224]. Similar to our results, a serum protein classifier was established that identified patients with poor [median OS, 7.3 months; n = 60 (45%)] and good [median OS, not reached; n = 73 (55%)] outcomes. This classifier was shown in blinded validation to be predictive for atezolizumab versus docetaxel for OS and PFS ($P_{interaction} < 0.01$). In that study, as in our own, there was no association between test classification and tumor PD-L1 expression; there was also no association with TMB. Also, among the biological processes that were most significantly associated with classification by PSEA, acute inflammation and complement activation ranked in the top three.

There are some limitations to our results. Obviously, the number of patients was low and all three immunotherapy-treated cohorts came from one geographic area and were investigated retrospectively. Also, for historical reasons, validation blinded to all clinical data was only possible for validation set 2. Although we made strong efforts to obtain sufficient tumor tissue samples, we were not able to obtain PD-L1 expression data on all patients. Several factors contributed to this: many patients are diagnosed on the basis of cytology alone and so have no tissue available for PD-L1 analysis; at the time of treatment initiation for these patients, use of PD-L1 expression was still somewhat investigational; and positive PD-L1 expression status was not mandatory for administration of nivolumab in the second- and higher-line setting. Unfortunately, TMB data were not collected. Investigation of larger cohorts with more complete information on TMB and PD-L1 expression would be useful to examine with more precision the level of association of these markers and how much complementary information each can provide to predict outcome. The non-immunotherapy-treated control set was small and restricted to one therapy. It would be of interest to study the performance of the test in larger control cohorts in other standard-of-care non-immunotherapy regimens to be able to explore the test's predictive potential.

Additional validation of the test in other larger cohorts of patients treated with CPIs is necessary. So far, we have investigated the ability of the test to stratify outcome for patients receiving checkpoint blockade monotherapy in the second- and higher-line setting, after platinum-based chemotherapy. However, now immunotherapy is moving into the first-line setting, either as monotherapy for patients with PD-L1 expression greater than 50%, or in combination with chemotherapy. It is of interest to evaluate the performance of the test in these first-line settings. A prospective trial, comparing outcomes between mono-immunotherapy and the chemo-immunotherapy combination in first-line patients with high PD-L1 expression is in the final stages of design. Studies in earlier stage patients receiving durvalumab with chemoradiation would also be informative. Evaluation of the test with appropriate comparator non-immunotherapy regimens in a prospective, randomized setting would be required to unequivocally determine its predictive power and clinical utility.

Supplemental material

The supplemental material can be found at:

Incluced content in this thesis:
Supplemental results:
Supplementary Table 9: Classification concordance (sensitive vs intermediate vs resistant)
Supplementary Table 10: Classification concordance (resistant vs not resistant)
Supplementary Table 11: Classification concordance (sensitive vs not sensitive)
Supplementary Table 12: Response by test classification in the development cohort
Supplementary Table 13: Multivariate Analysis of OS and PFS resistant vs not resistant
Supplementary Table 14: Multivariate Analysis of OS and PFS sensitive vs not sensitive
Supplementary Table 19: PD-L1 status by test classification (pooled second line patients)
Supplementary Table 21: PSEA for association of biological processes
Supplementary Figure 5: Kaplan-Meier plots of OS by test classification
Supplementary Figure 8: Dot plot of PD-L1 staining by test classification
Supplementary Figure 9: Kaplan-Meier plots of OS by test classification

Supplemental Data: Results

5. Reproducibility of Test Classifications

To assess the reproducibility of the test, the test was run from scratch on samples from Validation Set 1 on two occasions more than one year apart. The classifications obtained for Run1 and Run2 are compared in Supplementary Table 9 for classification sensitive vs intermediate vs resistant, in Supplementary Table 10 for the binary combination resistant vs not resistant (intermediate and sensitive), and in Supplementary Table 11 for the binary combination not sensitive (intermediate and resistant) vs sensitive. Classification concordance is 85% for the three-way classifications, 91% for the resistant / not resistant combination and 93% for the not sensitive / sensitive combination.

Supplementary Table 9 - Classification concordance (sensitive vs intermediate vs resistant)

		Run2		
		resistant (N=37)	**intermediate (N=30)**	**sensitive (N=31)**
Run1	**resistant (N=40)**	34	6	0
	intermediate (N=22)	2	19	1
	sensitive (N=36)	1	5	30

Supplementary Table 10 - Classification concordance (resistant vs not resistant)

		Run2	
		resistant (N=37)	**not resistant (N=61)**
Run1	**resistant (N=40)**	34	6
	not resistant (N=58)	3	55

Supplementary Table 11 - Classification concordance (sensitive vs not sensitive)

		Run2	
		not sensitive (N=67)	**Sensitive (N=31)**
Run1	**not sensitive (N=62)**	61	1
	sensitive (N=36)	6	30

1. Association of response with test classification

Supplementary Table 12 - Response by test classification in the development cohort

n (%)		resistant (N=41)	intermediate (N=43)	sensitive (N=32)
		n (%)	n (%)	
Response	CR	0 (0)	0 (0)	1 (3)
	PR	4 (10)	4 (9)	8 (25)
	SD	1 (2)	11 (26)	7 (22)
	PD	35 (85)	21 (49)	9 (28)
	NA	1 (2)	7 (16)	(22)

2. Multivariate Analysis of Development Cohort

Supplementary Table 13 - Multivariate Analysis of OS and PFS for development cohort by test classification resistant vs not resistant

	OS		PFS	
	HR (95% CI)	p value	HR (95% CI)	p value
Test classification (not resistant vs resistant)	0.59 (0.34-1.03)	0.062	0.53 (0.32-0.89)	0.015
ECOG PS (1 vs 0)	1.71 (0.90-3.22)	0.100	1.36 (0.78-2.35)	0.277
ECOG PS (≥2 vs 0)	4.67 (2.05-10.66)	<0.001	2.50 (1.19-5.25)	0.016
Never vs ever smoker	1.88 (0.84-4.23)	0.126	1.20 (0.54-2.65)	0.657
Squamous vs Non-squamous	1.02 (0.56-1.84)	0.960	1.04 (0.62-1.76)	0.876
PD-L1 (<1% vs ≥ 1%)	1.53 (0.79-2.95)	0.205	1.40 (0.76-2.58)	0.285
PD-L1 (NA vs ≥ 1%)	0.85 (0.41-1.77)	0.669	0.86 (0.45-1.65)	0.655

Supplementary Table 14 - Multivariate Analysis of OS and PFS for development cohort by test classification sensitive vs not sensitive

	OS		PFS	
	HR (95% CI)	p value	HR (95% CI)	p value
Test classification (sensitive vs not sensitive)	0.60 (0.30-1.20)	0.150	0.63 (0.35-1.15)	0.132
ECOG PS (1 vs 0)	1.64 (0.86-3.13)	0.133	1.39 (0.80-2.40)	0.245
ECOG PS (≥2 vs 0)	4.80 (2.11-10.91)	<0.001	2.68 (1.28-5.57)	0.009
Never vs ever smoker	2.14 (0.92-4.95)	0.077	1.31 (0.58-2.95)	0.512
Squamous vs Non-squamous	0.99 (0.55-1.81)	0.988	1.05 (0.62-1.78)	0.856
PD-L1 (<1% vs ≥ 1%)	1.65 (0.86-3.14)	0.130	1.66 (0.92-2.99)	0.091
PD-L1 (NA vs ≥ 1%)	1.02 (0.50-2.07)	0.958	1.08 (0.57-2.04)	0.811

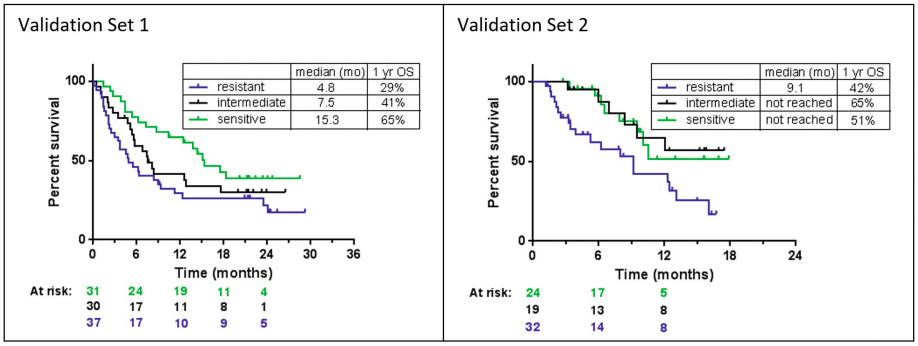
Supplementary Table 19 - PD-L1 status by test classification (pooled second line patients)

	resistant (N=96)	intermediate (N=80)	sensitive (N=73)
PD-L1 Positive (≥1%)	21 (22)	20 (25)	16 (22)
PD-L1 Negative (< 1%)	25 (26)	28 (35)	16 (22)
NA	50 (52)	32 (40)	41 (56)

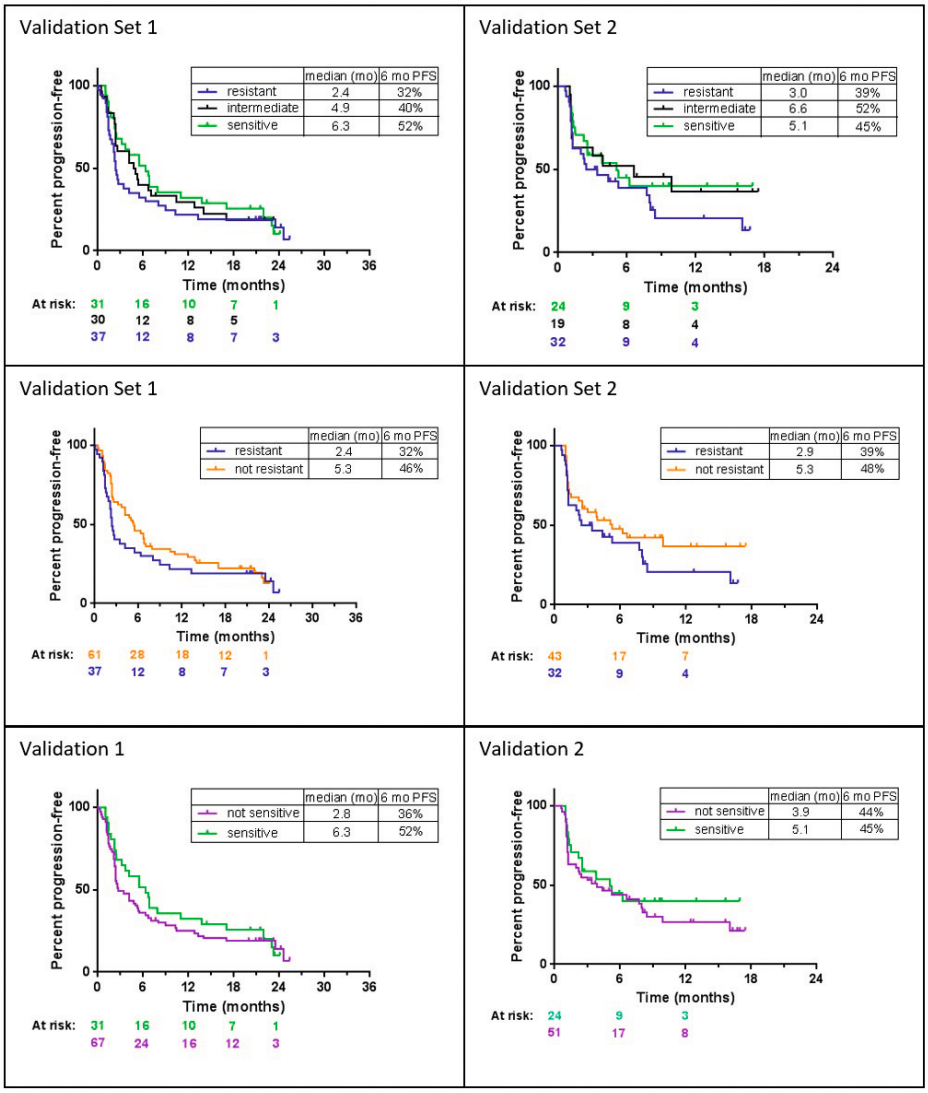
NA=not available

Supplementary Table 21 - PSEA for association of biological processes with test classification sensitive vs not sensitive

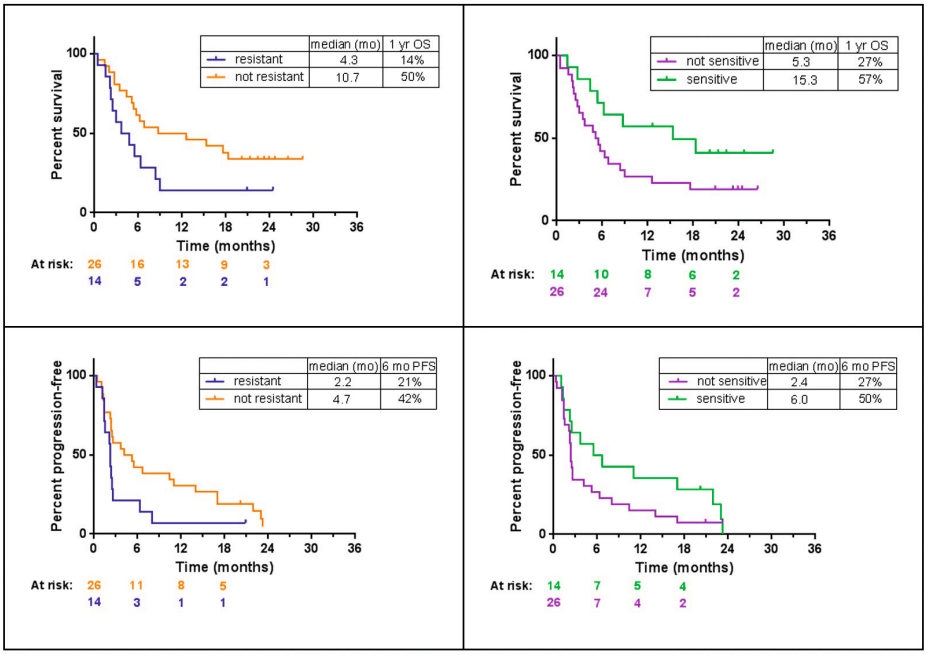
Biological Process	p value of association	FDR
Immune tolerance and suppression	0.0035	<0.1
Acute inflammatory response	0.0154	<0.2
Acute phase response	0.0170	<0.2
Cytokine production involved in immune response	0.0665	<0.4
Complement activation	0.1372	<0.5
Innate immune response	0.1523	<0.5
Angiogenesis	0.1532	<0.5
NK cell mediated immunity	0.1717	<0.5
B cell mediated immunity	0.2683	<0.7
Wound healing	0.2760	<0.7
Type2 immune response	0.3887	<0.8
Extra cellular matrix	0.4302	<0.8
Epithelial-Mesenchymal Transition	0.4332	<0.8
Chronic Inflammatory response	0.5087	<0.8
IFN type 1	0.5488	<0.8
IFN-Gamma	0.5558	<0.8
Type17 immune response	0.5576	<0.8
Response to hypoxia	0.5601	<0.8
Cellular component morphogenesis	0.6322	<0.8
T cell mediated immunity	0.6769	<0.8
Type 1 immune response	0.7802	<0.9
Glycolysis and positive regulators	0.8013	<0.9
Behavior	0.8487	<0.9



Supplementary Figure 5 - Kaplan-Meier plots of OS by test classification sensitive vs intermediate vs resistant in the Validation Sets



Supplementary Figure 8 - Dot plot of PD-L1 staining by test classification in the pooled analysis of second line patients with known staining. Less than 1% is shown as 0. Whiskers show minimum and maximum. Boxes show the median and quartiles. Median and first quartile are both 0% for resistant and intermediate. Median and first quartile are 0% and 0.5% for sensitive.



Supplementary Figure 9 - Kaplan-Meier plots of OS by test classification sensitive vs intermediate vs resistant in the chemotherapy set