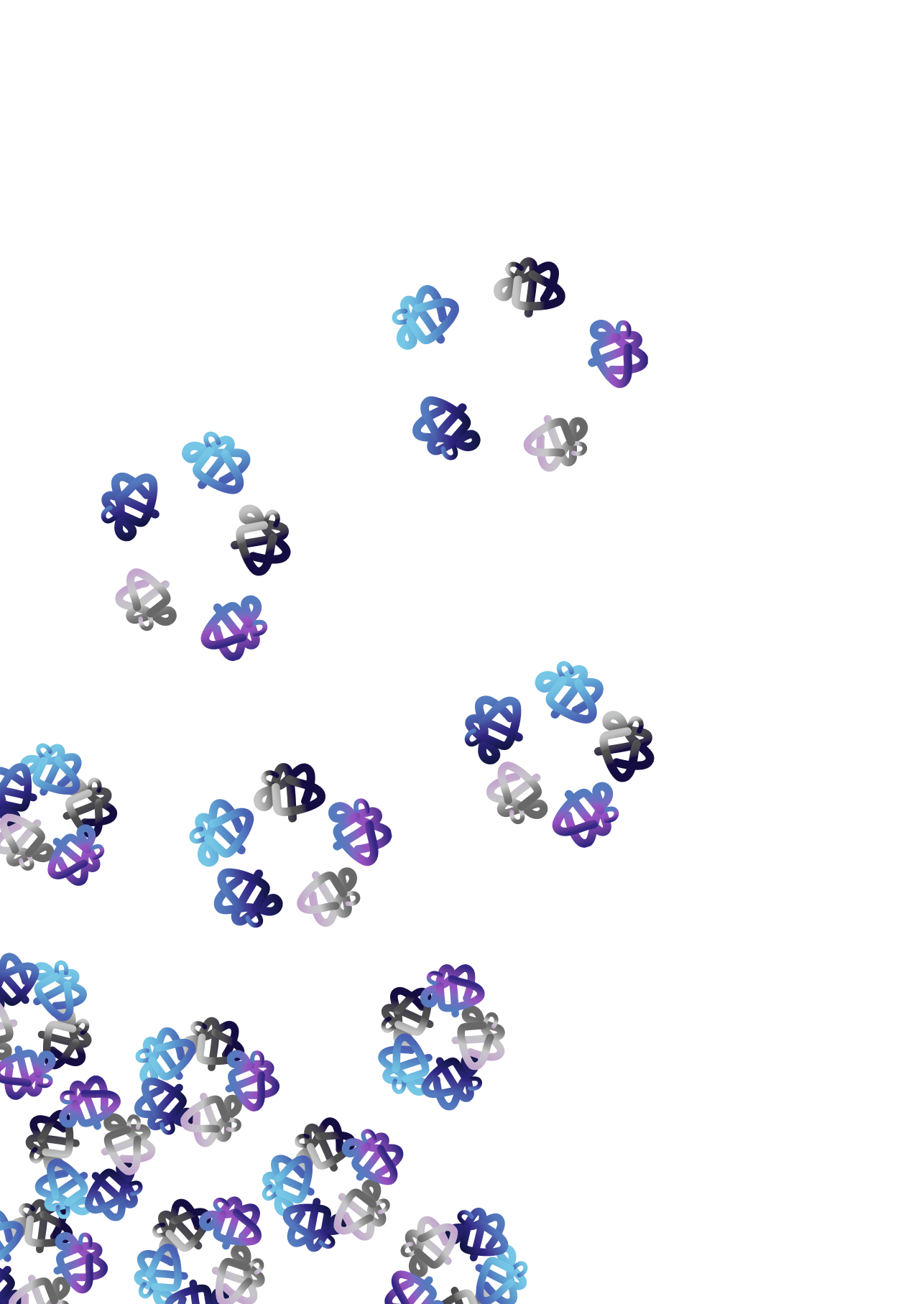
Biomarkers for the response to immunotherapy in patients with non-small cell lung cancer

Muller, M.



8 Validation of a clinical blood-based decision aid to guide immunotherapy treatment in patients with non-small cell lung cancer

M. Muller, R. Hoogendoorn, R. Moritz, V. van der Noort, M. Lanfermeijer, Catharina M. Korse, D. van den Broek, J. J. den Hoeve, P. Baas, H.H. van Rossum, M. M. van den Heuvel

Tumour biology: the journal of the International Society for Oncodevelopmental Biology and Medicine. 2021;43(1):115-27.

Abstract

Background: The widespread introduction of immunotherapy in patients with advanced non-small cell lung cancer (NSCLC) has led to durable responses but still many patients fail and are treated beyond progression.
Objective: This study investigated whether readily available blood-based tumor biomarkers allow accurate detection of early non-responsiveness, allowing a timely switch of therapy and cost reduction.
Methods: In a prospective, observational study in patients with NSCLC treated with nivolumab or pembrolizumab, five serum tumor markers were measured at baseline and every other week. Six months disease control as determined by RECIST was used as a measure of clinical response. Patients with a disease control < 6 months were deemed non-responsive. For every separate tumor marker a criterion for predicting of non-response was developed. Each marker test was defined as positive (predictive of non-response) if the value of that tumor marker increased at least 50% from the value at baseline and above a marker dependent minimum value to be determined. Also, tests based on combination of multiple markers were designed. Specificity and sensitivity for predicting non-response was calculated and results were validated in an independent cohort. The target specificity of the test for detecting non-response was set at > 95%, in order to allow its safe use for treatment decisions.
Results: A total of 376 patients (training cohort: 180, validation cohort: 196) were included in our analysis. Results for the specificity of the single marker tests in the validation set were CEA: 98·3% (95%CI: 90·9–100%), NSE: 96·5% (95%CI: 87·9–99·6%), SCC: 96·5% (95%CI: 88·1–99·6%), Cyfra21·1 : 91.8% (95%CI: 81·9–97·3%), and CA125 : 86·0% (95%CI: 74·2–93·7%). A test based on the combination of Cyfra21.1, CEA and NSE accurately predicted non-response in 32.3% (95% CI 22.6–43.1%) of patients 6 weeks after start of immunotherapy. Survival analysis showed a significant difference between predicted responders (Median PFS: 237 days (95%CI 184–289 days)) and non-responders (Median PFS: 58 days (95%CI 46–70 days)) ($p < 0.001$).
Conclusions: Serum tumor marker based tests can be used for accurate detection of non-response in NSCLC, thereby allowing early and safe discontinuation of immunotherapy in a significant subset of patients.

Keywords: Serum tumor marker, CEA, Cyfra, SCC, NSE, CA125, Nivolumab, response, Longitudinal

1. Introduction

Immune checkpoint based therapies for lung cancer have changed the therapeutic landscape of and survival from non-small cell lung cancer (NSCLC) [3, 5, 14]. Unfortunately, still a limited number of patients respond to immune checkpoint based treatment and non-responsiveness remains a clinical challenge [3, 24]. Therefore, treatment monitoring in order to detect non-responsiveness is of key importance and rapid detection of non-responsiveness potentially allows a prompt next in line treatment initiation avoiding unnecessary side effects and costs.

For NSCLC follow-up several circulating tumor biomarker are available [280-282]. Most biomarkers available in clinical practice have not been validated as monitoring tools [281, 283] . Tumor markers readily available at medical laboratories and potentially useful to monitor NSCLC treatment response include CA125, carcinoembryonic antigen (CEA), cytokeratin 19 fragments (Cyfra 21·1), neuron-specific enolase (NSE), and squamous cell carcinoma antigen (SCC) [284-286]. Though evidence supports the clinical application of some of these tumor biomarkers for lung cancer, no clear guidance is available [280, 282, 284, 287, 288]. The interpretation of these tumor biomarkers, when used for monitoring a specific cancer treatment, is therefore generally based on expert opinion and personal experience.

Recently, a method and software package, called ReMarker, was developed to assess the applicability of tumor marker changes after start of treatment in the response assessment [289]. We used this application to design and validate biomarker-response based tests that allow an accurate and early detection of non-responsiveness to immunotherapy for patients with NSCLC. This would allow early discontinuation of ineffective therapy and provide a window of opportunity for initation of subsequent other treatment opportunities. Besides, it would reduce potential side effects and costs. Our aim was to define and clinically validate an early response tool that accurately predicts non-response and can be easily applied in daily clinical practice based on changes in tumor markers during therapy.

2. Methods

2.1. Study population

In a prospective, observational study, patients with NSCLC treated with nivolumab or pembrolizumab were included. Serum tumor markers CA125, CEA, Cyfra 21·1, NSE, and SCC were measured. Using the data on clinical outcome a test for every separate tumor marker was designed: our aim was to design and optimize a test that identifies non-responders (as determined by RECIST at six months) as early as six weeks after starting treatment based on their serum marker values. Each such test was defined as positive if the value of that tumor marker met two criteria: (i) elevation of 50% compared to baseline and (ii) above a minimum value. A training cohort was used to determine the optimal minimum value for each tumor marker. An independent validation cohort was used to validate the resulting tumor marker tests. Also the performance of combining the results of the individual tumor marker tests were evaluated. In this single-center study all patients with NSCLC who started their treatment between March 2013 and September 2018 in The Netherlands Cancer Institute were included. Follow-up was available until

January 2019. All consecutive patients receiving immunotherapy in a variety of settings, such as routine care, early access, compassionate use program, and clinical trials, were treated according to corresponding protocols. Patient criteria for receiving nivolumab treatment were previously described [24] and can also be found in the supplemental material, as are the pembrolizumab criteria. If a patient had received immunotherapy in two different treatment lines, the initial treatment line was taken. Tumor markers were measured at baseline and prior to each consecutive cycle together with other routine blood assessment tests as standard of care. The monitoring of response was done with a CT scan before start of treatment, and after 6 weeks, 3 months and every 3 months thereafter. Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 were used, accordingly progressive disease (PD), stable disease (SD), and partial response (PR) [105]. Patients who were progressive before the endpoint of six months, were classified as having no clinical benefit (NCB), as previous described in Rizvi et al. [112]). Our study was approved by the local medical ethical committee (PTC NKI-AvL, NL45524.031.13), patient privacy committee and performed according to the institutional patient privacy protocols. In January 2017 all patients who had been treated with nivolumab at that time were randomly assigned to the training or validation cohort in a 2:1 ratio (Figure 1), as described in the sample size calculation (supplemental material). The training cohort was used to make and refine the ReMarker application (see below). After this randomization, no more patients were added to or removed from this training cohort. The validation cohort consisted of patients who were initially randomized to the validation cohort, and those who started their nivolumab treatment after January 2017 or who were treated with pembrolizumab, up to a 1:1 ratio.

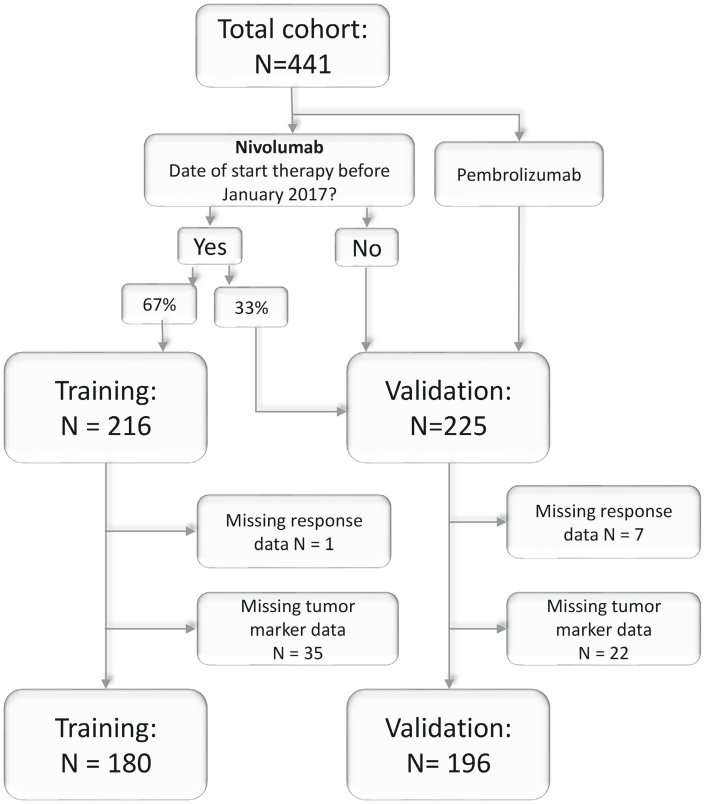


Figure 1 - Consort chart.
All patients treated with immunotherapy as second or higher line in the training and validation cohort.

2.2. Design of tumor marker test

Analysis of the obtained serum samples were performed on a daily (CA125,CEA, Cyfra 21·1, and NSE) or twice weekly (SCC) basis. CA125, CEA, Cyfra 21·1, and NSE were measured using a Cobas 6000 system (Roche diagnostics) and SCC was measured on a Kryptor system (Thermo Fisher), both according to the manufacturer's instructions. The applied reference ranges for the tumor markers were < 20 U/ml for CA125 (< 35 U/ml for premenopausal females), < 6g/L for CEA, < 1·9g/L for Cyfra 21·1, < 12·5g/L for NSE and < 2·0g/L and < 1·5g/L for SCC for males and females, respectively. The application ReMarker was used to study multiple time points and multiple cut-offs. The correlation to clinical response was visualized in Biomarker Response Characteristic plots (BReC plots) (Fig. S1) [289]. The baseline measurement was defined as minus 3 weeks until 0 weeks before start of treatment. A follow-up time point of 6 weeks was designated as primary optimization follow-up time, since in our practice this is the first clinical evaluation moment for response evaluation. This follow-up time point was defined as a measurement 5 or 6 weeks (±3 days) after start of treatment. If there was more than one

measurement in one of these periods, the latest measurement was taken. The training set was used to optimize the test per single tumor marker for the prediction of non-response, which was defined as PD, NCB or deceased after six months of immunotherapy treatment. The other patients were classified as responders. The following factors were taken into consideration for the design (and are also explained in Table S1): (I + II) In order to obtain an easy-to-calculate test, we defined our test as positive (i.e. predictive for non-response) when the marker increased with 50% from baseline and was above the marker dependent minimum value (Fig. S1); (III) The minimum value criterion was applied to exclude patients with small biomarker increases at low concentrations that results in large relative increases thereby reducing the effect of (pre-) analytical and biological "noise"; (IV) The optimal minimum value per marker was determined by calculating the specificity and sensitivity (Fig. S2); (V) Minimum values yielding a specificity of ≥ 97·5% in the training set per individual markers were considered a good cut-off; (VI) Minimum values yielding a sensitivity of > 20% were considered a good cut-off. An overview of the considerations can be found in Table S1. For each tumor marker we chose a minimum value satisfying the criteria in the training cohort (Fig. S2). Then, in the validation cohort, the sensitivity, specificity, positive predicted value (PPV) and negative predicted value (NPV), all with a 95% confidence interval, were calculated for the resulting test per tumor marker. After the best test per single tumor marker had been determined, the combination of tumor markers was tested in the training cohort (Table S2). A test was considered positive if at least one of the tumor markers increased with 50% above baseline. Only the tests in the training cohort that fulfilled abovementioned criteria were validated in the validation cohort, again in terms of sensitivity, specificity, NPV, and PPV, all with a 95% confidence interval. The performance of the tests was also investigated from week 2 until week 20, with biweekly tests, for both the training cohort and validation cohort in order to allow a more general application. Survival analyses and cox-regression analyses were performed assessing the predictive value of the tests for overall survival (OS) and progression free survival (PFS). OS was defined as the number of days between the day of start of treatment and date of death, PFS as the number days between the day of start of treatment and date of progression or death, whichever came first. SPSS (v25; SPSS, Chicago, USA) was used for the descriptive statistics. Descriptive statistics were expressed as mean ± SD if data were normally distributed and as median (interquartile range) if data were non-normally distributed. Between group comparisons were performed using Mann-Whitney U tests, two sample unpaired t-tests of Chi-Squared tests. From all the patients with a false-positive result, the medical record was checked for possible confounders.

Furthermore, a small cohort with patients who were treated with pembrolizumab in first line (rather than second line) was also available for analysis.

3. Results

3.1. Patients

A total of 441 patients were included in our study, 216 in the training set and 225 in the validation set (Table 1). From these patients, 389 patients were treated with nivolumab and 52 with pembrolizumab. A total of 65 patients were excluded from our analysis due to missing data (Fig. 1). The training cohort consisted of who 53 responders and 127 non-responders. (Table S3). The validation cohort consisted of 69 responders and 127 non-responders. There was a significant difference between the responders and non-responders with regard to the PD-L1 status ($p < 0{\cdot}001$).

3.2. Test design

The following test optimization minimum values were established for a test at 6 weeks: CA125 : 65 U/ml, CEA: 6g/L, Cyfra 21·1 : 4g/L, NSE: 20g/L, and SCC: 3·5g/L (Fig. S2). In the validation set, the specificity of CEA, NSE, SCC, Cyfra 21·1, and CA125 was 98·3% (95%CI: 90·9–100%), 96·5% (95%CI: 87·9–99·6%), 96·5% (95%CI: 88·1–99·6%), 91·8% (95%CI: 81·9–97·3%), and CA125 86·0% (95%CI: 74·2–93·7%) respectively (Table 2). Only the markers NSE and SCC showed a sensitivity below 20%. For SCC however, a small subset of patients with a squamous cell carcinoma showed an increase in the sensitivity of our test, without loss of specificity, from 6% (3·0–11·1%) to 15·4% (6·4 – 31·2%) (Table S6). The test accuracy for tumor marker to predict non-response was comparable between week 2 and 20 (Fig. 2).

Table 1 - Patient characteristics of the full cohort. All patients described have at least one baseline measurement and one follow-up measurement between weeks 2-20.

	TRAINING			VALIDATION				
	Non-responders (PD)	Responders (PR & SD)		Non-responders (PD)	Responders (PR & SD)		Total	
	N=127	N=53	p-value	N=127	N=69	p-value	N=376	p-value
Patient								
Male sex – no. (%)	75 (59.1)	27(50.9)	0.317	65(51.1)	35(50.7)	0.951	202(53.7)	0.273
Age (years) – mean (SD)	62.8(SD: 10.727)	64.3(SD: 8.190)	0.375	62.9 (SD: 8.9)	62.1 (SD: 8.9)	0.513	62.9 (SD: 9.5)	0.542
Smoking – no. (%)	25(19.7)	3(5.7)	0.010	17(13.4)	4(5.8)	0.249	49(13.0)	0.384
Pack years – mean (SD)	31.6(SD: 18.85)	35.5(SD: 19.3)	0.256	36.3 (SD: 19.5)	34.2 (SD: 17.5)	0.486	34.3 (SD 18.9)	0.240
WHO ≥ 2– no.(%)	20(15.7)	3(5.7)	0.047	13(10.2)	3(4.3)	0.156	39(10.4)	0.148
Tumor characteristics								
Pathology– no.(%)								
Adenocarcinoma	94(74.0)	33(62.3)		84(66.1)	39(56.5)		250(66.5)	
Squamous	22(17.3)	13(24.5)	0.286	24(18.9)	18(26.0)	0.285	77(20.5)	0.176
Other	11(8.7)	7(13.2)		19	12(17.4)		49(13.0)	
Mutations– no.(%)								
EGFR positive	4(3.1)	0	0.182	7(15.0)	2(2.0)	0.297	13(3.5)	0.192
KRAS positive	37(29.1)	15(28.3)	0.827	41(32.3)	25(36.2)	0.923	118(31.4)	0.020
BRAF	5(3.9)	2(1.6)	0.813	4(3.1)	1(1.4)	0.440	12(3.2)	0.405
ALK	0	0	-	2(1.6)	1(1.4)	0.898	3(0.8)	0.053
PD-L1 – no.(%) [1]								
Unknown	65(51.2)	20(37.7)	0.358	55(43.3)	28(40.6)	0.022	167 (44.4)	0.129
PD-L1 <1%	38(61.3)	17(51.5)	0.025	40(55.6)	14 (33.3)	<0.001	109 (52.2)	<0.001
PD-L1 >1%	24(38.7)	16(48.5)		32(44.4)	28 (66.7)		100 (47.8)	0.536
PD-L1 >50%	8(12.9)	10(18.9)	-	17(23.6)	24(57.1)		59 (28.2)	
Brain Metastasis – no.(%)	26(20.5)	12(22.6)	0.745	25(19.7)	11(15.9)	0.578	74 (19.6)	0.536

Table 1 - Patient characteristics of the full cohort. All patients described have at least one baseline measurement and one follow-up measurement between weeks 2-20. *(Continued)*

	TRAINING			VALIDATION				
	Non-responders (PD)	Responders (PR & SD)		Non-responders (PD)	Responders (PR & SD)		Total	
	N=127	N=53	p-value	N=127	N=69	p-value	N=376	p-value
Treatment								
Nivolumab	127(100)	53(100)	-	104(81.9)	51(73.9)	0.190	335 (89.1)	<0.001
Pembrolizumab	0	0	0.465	23(18.1)	18 (26.1)	0.400	41 (10.9)	
Line of treatment – no.(%)								
1st line	3(2.4)	0		2(1.6)	3(4.3)		8 (2.1)	
2nd line	90(70.9)	40(75.5)	-	99(78.0)	54(78.3)	-	283(75.3)	0.304
≥ 2nd line	33(26.9)	12(22.6)		26(20.5)	11(15.9)		82(21.8)	
Deceased after 6 months	74(58.3)	0		59(46.5)	0		133(35.4)	0.026
Comorbidities								
Auto Immune Disease – no.(%)	6(4.7)	0	0.106	8(6.3)	6(8.7)	0.538	20 (5.3)	0.119

[1] *Percentages shown are based on total known PD-L1 scores.*
Abbreviations: N: Number of patients; SD: Standard Deviation; no.: Number of patients, ECOG performance-status score: European Cooperative Oncology Group performance status score, this is a score ranging from 0 to 5, where 0 indicates no symptom, 1 indicates mild symptoms and above 1 indicates greater disability; EGFR: Epidermal Growth Factor Receptor; KRAS: Kirsten rat sarcoma viral oncogene; BRAF: v-raf murine sarcoma viral oncogene homolog B *; ALK: Anaplastic Lymphoma Kinase; PD-L1: Programmed death ligand 1.*

Table 2 - Results of the training cohort and the tests validated in the validation cohort.

Test	Minimum value	Results training			Results validation		
Single marker test							
		Sensitivity	**Specificity**	**PPV**	**Sensitivity**	**Specificity**	**PPV**
CA125	**65 U/ml**	21.7% (13.4-32.1%)	97.9% (88.7-100%)	94.7% (74.0-99.9%)	25.0% (16.0-35.9%)	86.0% (74.2-93.7%)	71.4% (51.3-86.8%)
CEA	**6 µg/L**	19.8% (12.0-29.8%)	100% (92.8-100%)	100% (80.5 – 100%)	20.7% (12.6-31.1%)	98.3% (90.9-100%)	94.4 % (72.7 – 99.9%)
	12 µg/L	15.1% (8.3-24.5%)	100% (92.8-100%)	100% (75.3-100%)	16.1% (9.1-25.5%)	98.4% (91.2-100%)	93.3% (68.1-99.8%)
Cyfra 21.1	**4 µg/L**	31.8% (22.1-42.8%)	100% (92.8-100%)	100% (87.2-100%)	34.5% (24.6-45.5%)	91.8% (81.9-97.3%)	85.7% (69.7 – 95.2%)
	8 µg/L	23.5% (15.0-34.0%)	100% (92.8-100%)	100% (83.2-100%)	25.3% (16.6-35.8%)	95.1% (86.3-99.0%)	88.0% (68.8-97.5%)
NSE	**20 µg/L**	13.3% (6.8-22.5%)	100% (92.1-100%)	100% (71.5-100%)	12.7% (6.2-22.1%)	96.5% (87.9-99.6%)	83.3% (51.6 – 97.9%)
	40 µg/L	4.8% (1.3-11.9%)	100% (92.1-100%)	100% (39.8-100%)	8.9% (3.6-17.4%)	96.5% (88.0-99.6%)	77.8% (40.0-97.2%)
SCC	**3.5 µg/L**	9.6% (4.3-18.1%)	97.9% (88.9-100%)	88.9 (51.8 – 99.7%)	2.4% (0.3-8.5%)	96.5% (88.1-99.6%)	50% (6.7-93%)
Combinations							
		Sensitivity	**Specificity**	**PPV**	**Sensitivity**	**Specificity**	**PPV**
Cyfra 21.1 OR CEA	**Cyfra: 4µg/L CEA: 6 µg/L**	38.4% (28.1-49.5%)	100% (92.8-100%)	100% (89.4-100%)	40.2% (29.9-51.3%)	91.8% (81.9-97.3%)	81.8% (64.5 – 93.0%)
Cyfra 21.1 OR CEA OR NSE	**Cyfra: 4µg/L CEA: 6 µg/L NSE: 20 µg/L**	38.4% (28.1-49.5%)	100% (92.8-100%)	100% (89.4-100%)	43.7% (33.1-54.7%)	91.9% (82.2-97.3%)	82.7% (66.3 – 93.4%)
Cyfra 21.1 OR CEA	**Cyfra: 8 µg/L CEA: 12 µg/L**	30.2% (20.8-41.1%)	100% (92.8-100%)	100% (86.8-100%)	28.7 % (19.5 – 39.43%)	95.1% (86.3-99.0%)	89.3 % (71.8-97.8%)
Cyfra 21.1 OR CEA OR NSE	**Cyfra: 8 µg/L CEA: 12 µg/L NSE: 40 µg/L**	30.2% (20.8-41.1%)	100% (92.8-100%)	100% (86.8-100%)	32.2 % (22.6 – 43.1%)	95.2% (86.5 – 99.0%)	90.3 % (40.7 – 59.3%)

Each marker test was defined as positive for the prediction of non-response if the value of that tumor marker met two criteria: (i) elevation of 50% compared to baseline and (ii) above a minimum value(second column in the table). All results, such as sensitivity, are given as a percentage (95% confidence interval). PVV: positive predicted value. µg/L: microgram per liter; U/ml: Units per milliliter.

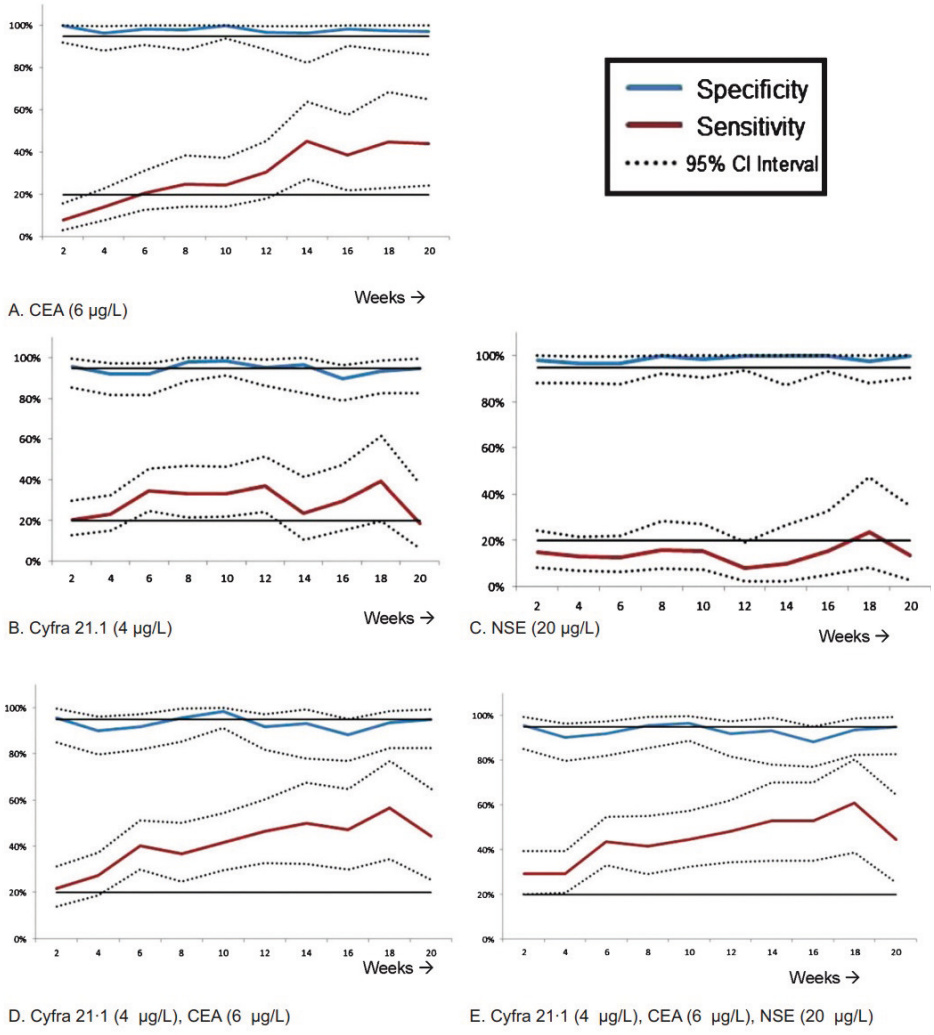


Figure 2 - Test characteristics for week 2–20 in the validation cohort, shown as sensitivity and specificity per week.
The horizontal as indicates the tests done every other week. Every time point displayed is that week and the week before (i.e. the time period for week 2 is week 1-2). If there was more than one measurement in this time period, the latest measurement was taken. The combination of markers were considered positive if at least one of the tumor markers had a positive test result. The two, straight lines indicate 20% and 95% respectively and are choosen for improved visibility. *µg*/L: microgram per liter; U/ml: Units per milliliter.

In the validation set the combination of Cyfra 21.1, CEA with or without NSE showed a specificity of 91.9% and 91.8% respectively and a sensitivity of 40.2–43.7% (Table 2). With these results, we decided to also validate a more stringent test by doubling the minimum value (Table S2). The specificity increased to 95.1% (86.3– 99%) with NSE and 95.2% (86.5–99.0%) without NSE at the cost of a lower sensitivity (28.7–32.2%). The results of the 23 performed tests in the training set with different markers can be found in the

supplemental material (Table S2 and Fig. S5). Also for the combination tests, we studied the same minimum value in other serial time points during treatment (week 2– week 20). The diagnostic performance of combined tumor biomarker tests for different follow-up are presented in Fig. 2.

3.3. False positive analysis

With the test at 6 weeks there were in total 13 patients (3.5% of the total cohort) with a false-positive result (Table S5). In the total cohort, 9 patients showed a false-positive result for CA125. In five out of these 13 patients Cyfra 21.1 showed a false-positive result. Two out of these five patients had a PR at 6 months from which one patient actually had a pseudo progression at 6 weeks, as was confirmed with a CT-scan. The other patient had an active hyperthyroidism at the start of treatment, which might explain the increases of the tumor markers (up to 7012% for SCC). For the other two patients, no specific explanation was found (TableS5)

3.4. Survival outcome

The median OS and PFS for the patients in the validation set were 363 days (95% CI 317–409 days) and 130 days (95%CI 98–162 days) respectively (Fig. 3). The median OS and PFS of patients depicted as non-responsive versus responsive were 153 days (95% CI 139–167) and 58 days (95%CI 46–70 days) versus 450 days (95% CI 347–553 days) and 237 days (95%CI 185–289) respectively ($p < 0.001$).

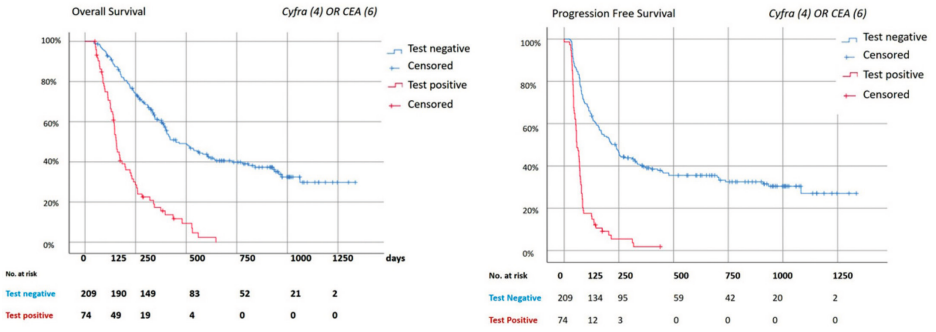


Figure 3 - Survival analysis.

Kaplan Meier analysis for the combination of Cyfra (4µg/L) and CEA (6µg/L). The combination of markers were considered positive if at least one of the tumor markers had a positive test result. All the analysis were done with the patients who had a test at week 6, as described in table S3. The median follow-up time was 322 days (IQR: 157–606 days). Date of last follow-up was 28th of January, 2019. A: Overall Survival. Median overall survival: 363 days (95% CI 317–409 days). Median OS negative test: 450 days (95% CI 347–553 days); Positive test: 153 days (95% CI 139–167 days). Log Rank (Mantel-Cox): $p < 0.001$. B: Progression Free Survival. Median progression free survival (PFS) 130 days (95% CI 98–162days). Median PFS negative test: 237 days (95% CI 185–289). Median PFS positive test: 58 days (95% CI 46–70 days). Log rank (Mantel-Cox) $p < 0.001$.

3.5. Pembrolizumab first line

In a small cohort of 31 patients who received pembrolizumab as first line treatment, an analysis was done. Results were comparable (Table S7 and S8).

4. Discussion

With the introduction of immunotherapy for metastasized NSCLC and its limited efficacy more tailored treatment strategies are needed. As far as we know, this is the first study that describes how to use liquid biopsy data for early treatment decisions in patients without clinical benefit from immunotherapy. In this prospective, observational study cohort a serum tumor markers panel was clinically validated as an early response tool that accurately predicts non-response to immunotherapy. These results indicate that serum tumor markers can be used to identify patients in which treatment can be discontinued early safely because it is ineffective. This potentially results in lower risk of side effect, lower costs, and allows alternative treatment options, while the patient is still in a good condition.

A commonly used and investigated liquid biopsy biomarker is ctDNA, which derives from normal physiological tissue remodeling events, necrosis and/or apoptosis of cancer cells [290-292]. The study of Goldberg et al. [250] showed the dynamics of ctDNA during immunotherapy treatment. In this study all patients with confirmed PR showed a ctDNA drop of > 50%, suggesting this is a helpful tool for monitoring response during treatment, although the dynamics of ctDNA in patients with progressive disease were more dynamic. Also, the strictly individual patterns of mutations complicate implementation in general practice of ctDNA-based response assessment and moreover a technical standardization for ctDNA is not yet available [292]. On the contrary, tumor markers are widely used, measured and implemented in clinical practice for years, making them a good alternative as a potential liquid biopsy. There is some literature about the role of serum tumor markers to assess efficacy of systemic treatment of NSCLC. Noonan et al. [293] showed in their study in patients with a targetable driver mutation in a smaller analysis that 59% of these patients, mostly responders, showed an increase right after start of their treatment. In the majority of patients marker concentration in plasma normalized to the baseline value during treatment. This shows the possible relation between tumor response and the measured markers. Furthermore, in the recent article of Dal Bello et al. [288], they measured CEA, Cyfra 21.1, and NSE at multiple time points. Their aim was to use these tumor markers for the monitoring of response (PR and SD). With their designed test, a decrease of 20%, identified responders. Interestingly enough, they also found that their test yielded similar results in the first and the fourth cycle of nivolumab. However, the study did not provide a tool to use these markers in optimizing treatment strategies, neither did other studies [282]. Therefore, our dataset with more than 400 patients and serial tumor marker data is contributing to the development of a clinical tool.

The requirements of a tumor marker test for early treatment decisions are depending on the clinical application. The current standard of care is to treat all patients with immunotherapy with or without chemotherapy, depending on the PD-L1 status [294]. A high specificity is required to prevent discontinuation of treatment in patients with a potential benefit. This approach, was also advocated in a study on the usage of an electronic nose. De Vries et al. [260] were able to identify 24% of the non-responders at baseline by exhaled breath analysis. On the other hand, the test should have an added value. Therefore the percentage of patients who will not respond and have a positive test, in other words sensitivity, must be contributing to current standards (e.g. radiological response and clinical assessment). In this study, we aimed to find the right balance of

these factors per individual and in combination of different tumor markers. Although not all individual markers showed a sufficient sensitivity (NSE and SCC), combining markers increases the sensitivity thereby optimizing its clinical utility.

Decisions during treatment are depending on radiological assessment and clinical performance; often treatment is continued despite the fact that the condition of the patient is deteriorating. In our cohort, the specificity of the CT-scan was 96.8% [24], but the therapy is only discontinued following confirmed radiological progression or in case of clinical deterioration. Besides, often no measurable lesions are available for radiological response assessment. Having established tumor markers as robust tool to establish non-responsiveness, we postulate that tumor markers can improve early treatment decision making. In the future, combining radiology and tumor markers, together with assessment of the clinical condition, will likely improve overall test accuracy.

Spikes are a well-known phenomenon seen in liquid biomarker research in patients with response [295, 296], which is also shown for immunotherapy [250]. However, in our study, spikes were not common (Fig. S6). Nevertheless, there were patients with a false-positive result. Renal failure, liver failure or (other) lung diseases are known causes for multiple different elevated tumor markers [297]. We did not see this in our cohorts, maybe due to the selection criteria of immunotherapy. What is more commonly known is the lack of accuracy of CA125. False positive results are often present in case of a serositis [298]. There was one patient with an active thyroiditis (Table S5A, Patient A) with extremely elevated tumor marker levels. We are not aware of data supporting the correlation of thyroiditis with tumor marker elevation. Our findings suggest that tumor marker tests should be treated with caution in case of an active thyroiditis.

A strength of this study is the homogeneous patient population with mainly ≥ 2nd line NSCLC patients treated with single agent immune checkpoint inhibitors and the use of independent training and validation cohort, which makes it an robust analysis. However, there are a few limitations of our study to be considered. Firstly, therapeutic options are rapidly changing and patients are currently treated with immunotherapy as their first line of treatment[293]. We included a small analysis with first-line pembrolizumab patients, in order to assess the utility of the tests in the current standard of treatment. Mature data of a small cohort (n = 23) of patients, treated with first line pembrolizumab, was available and results were comparable. However, larger validation studies are warranted. Secondly, there training cohort consisted of patients who were treated with nivolumab only. We are uncertain if this might cause bias. Pembrolizumab and nivolumab are both PD-1 inhibitors and the validation results in the pembrolizumab cohort were comparable. Thirdly, in our study, we validated a minimum value instead of a percentage. These minimum values and reference values differ between different hospitals. However, all of the chosen minimum values are more or less multiplied by a round number between one and three (Fig. S2), allowing a relatively easy validation of these tests in other hospitals.

All in all, in this study we designed and validated tests with single and multiple serum based tumor markers for the early prediction of non-response. Based on our results, serum tumor marker based response monitoring can be used for clinical decision making in NSCLC treated with immunotherapy. Future studies are required to determine the added value in clinical practice.

Supplemental data

The supplemental material can be found at:

Contents included in this thesis:
Figure S1 – Schematic view
Figure S2 – Cut-off check training set
Figure S3 – Week 2-20 for the training cohort
Figure S4 – Week 2-20 for the validation cohort
Table S1 – Considerations for the development of a marker test
Table S5 – False-positives for test at week 6.
Table S6 – Sub analysis Squamous Cell Carcinoma and SCC marker
Table S7 – Patient characteristics of the first line pembrolizumab cohort
Table S8 – Analyses first line pembrolizumab cohort

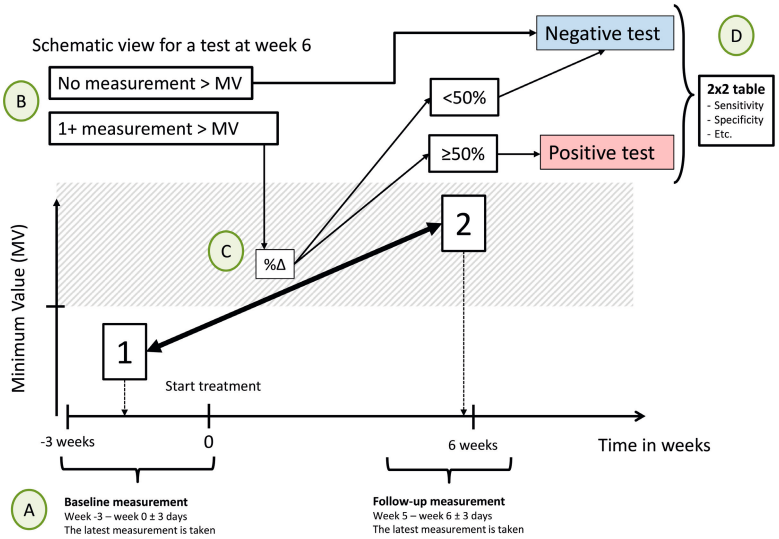


Figure S1 - Schematic view of the application
A: The application checks if a patient has measurements (in figure referred to as 1 and 2) within the given timeframes. If there is more than one measurement in a given baseline or follow-up timeframe, the last measurement will be used. If measurements within one of the two timeframes are missing, the results are obviously defined as non-conclusive (and therefore will not be add to the 2x2 cross table). **B:** The program checks if at least one of these measurement is above the minimum value (MV). The minimum value is a threshold used to avoid background noise from non-relevant increases of a marker. If both values are below this threshold, the test will be defined as negative. **C:** If abovementioned criteria are fulfilled, than the difference in percentage (%Δ = (measurement 2 – measurement1) / measurement1 x 100%) will be calculated**.** In our study, and in alignment with the results from the ReMarker analysis, the test was deemed to be positive when the marker in a patient had increased with 50%. **D:** The outcome of the test of a patient is used in a 2x2 cross table and compared to clinical data.

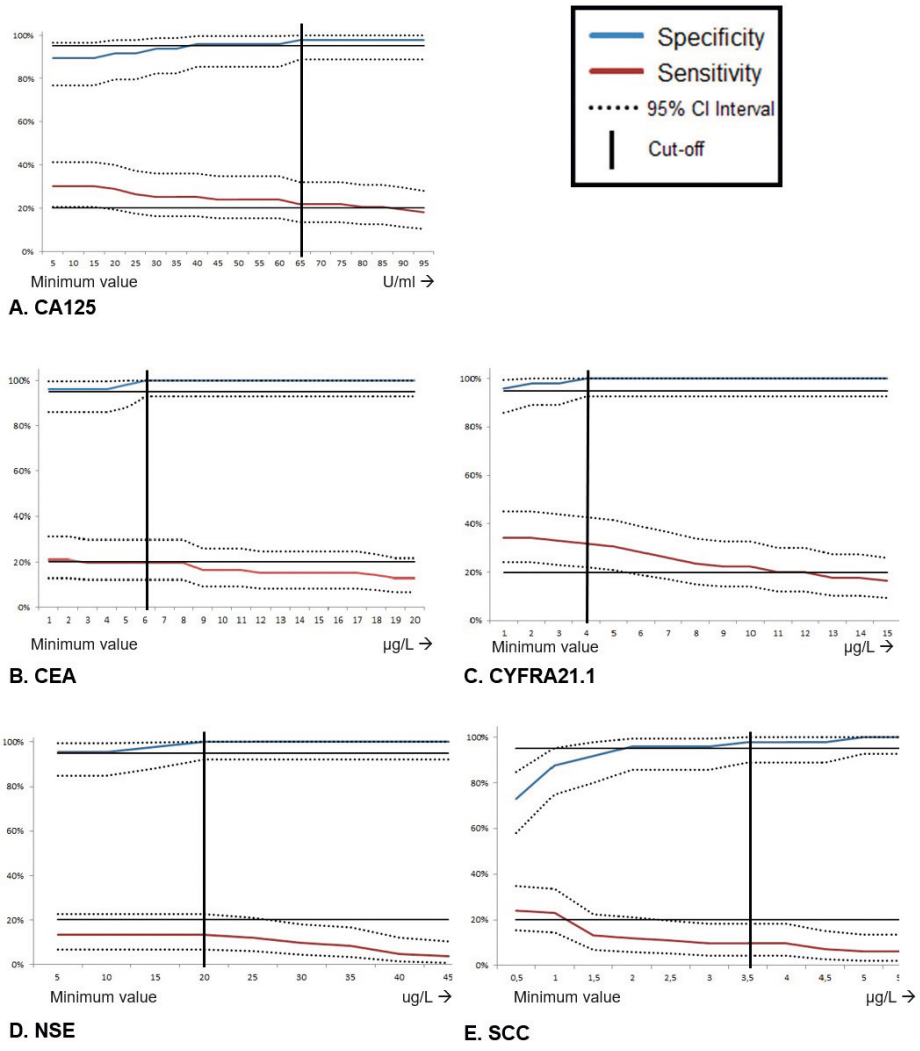


Figure S2 – The optimal minimum value at week 6 in the training cohort
µg/L: microgram per liter; U/ml: Units per milliliter.

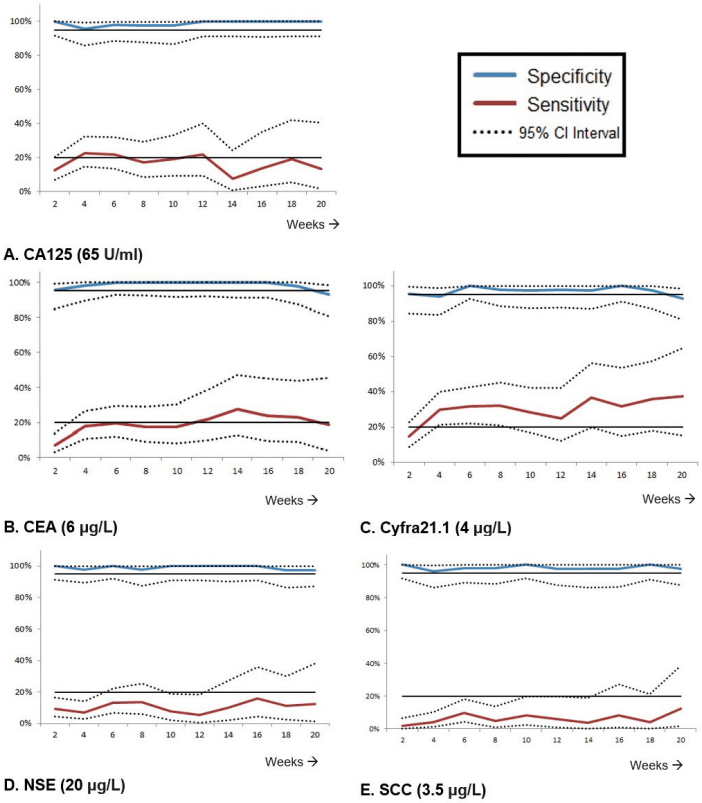


Figure S3 – Characteristics of the serum tumor marker cut-off values for week 2 – 20 in the training cohort, shown as specificity and sensitivity.

The horizontal axis indicates the tests done every other week. Every time point displayed is that week and the week before (i.e. the time period for week 2 is week 1-2). If there was more than one measurement in this time period, the latest measurement was taken. The two, straight lines indicate 20% and 95% respectively and are choosen for improved visibility.µg/L: microgram per liter; U/ml: Units per milliliter.

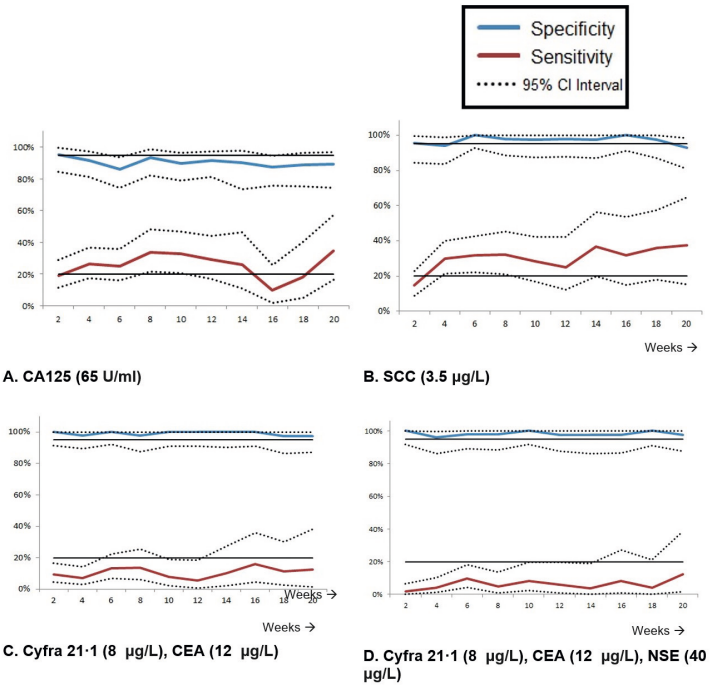


Figure S4 – Results of the test characteristics of tumor markers in the validation set, shown as sensitivity and specificity per week.

Notes: The combination of markers were considered positive if at least one of the tumor markers had a positive test result. The horizontal as indicates the tests done every other week. Every time point displayed is that week and the week before (i.e. the time period for week 2 is week 1-2). If there was more than one measurement in this time period, the latest measurement was taken. The two, straight lines indicate 20% and 95% respectively and are choosen for improved visibility. µg/L: microgram per liter; U/ml: Units per milliliter.

Table S1 - Considerations for the development of a marker test.

Considerations		Single marker test
1	Easy to use in clinical practice	
2	>50% increase	>50% increase compared to baseline Based on earlier BReC plot analysis [14] Considered easy-to-calculate
3	Minimum value	The minimum value is used instead of the reference value of a marker. Reference values in the clinical chemistry are based on the fact that 95% of all test results lay below the reference value, measured in the healthy population [294]. However, the aim is to identify non-responders in a group of lung cancer patients. Therefore, we introduced a new value, the minimum value. The role of the minimum value criterion is applied to exclude patients with small biomarker increases at low concentrations that results in large relative increases thereby reducing the effect of (pre-) analytical and biological "background noise" At least one of the measurements should be above the minimum value (as described in figure S1)
4	Optimal minimum value	The optimal minimum value per marker was determined by calculating the specificity and sensitivity for predicting the clinical endpoint per minimum value (figure S2)
5	Specificity > 97,5%	The test was developed as an early treatment decision tool and should be very accurate in detecting non-responsiveness (to safely discontinue treatment) Minimum values yielding a specificity of ≥97,5% in the training set per individual markers were considered a good cut-off, in order to increase the likelihood to achieve a specificity of >95% in the validation set
6	Sensitivity >20%	The test should have added value over the current standard, i.e. decisions based on radiological and clinical assessment. Observation in the training set is that about 20% of the patients discontinue treatment within 6 weeks based on radiological assessment and/or experienced clinical deterioration. Therefore a sensitivity >20% was considered a good cut-off. A combination of tumor markers was assumed to increase in sensitivity, therefore for the single biomarkers a sensitivity less than 20% was accepted in the training cohort
		A test with a combination of markers
7	Combination of single marker tests	Two single marker tests with the previous described characteristics are used of a combination test
8	At least one test is positive	The combination of markers were considered positive if at least one of the tumor markers had a positive test result, defined by an increase of the marker concentration according to abovementioned criteria When at least one of the markers is positive, the sensitivity of the test is likely to increase.
9	Two times the minimum value	Considered easy to calculate The combination of markers may lead to a decrease in specificity, since all patients with a false-positive tests are added together in the combination test, leading to more false-positive results.

Table S5.A - False positives in the final test (Cyfra/CEA/NSE)

ID	Diagnosis	Treatment	Response	Cohort	CA125	CEA	Cyfra	NSE	SCC	Explanation
A	Ad	Nivo	PR	Val	424%	225%	218%	89%	7012%	Hyperthyroidism / thyroiditis
B	Ad	Pembro	PR	Val	205%	-	61%	-	-	Pseudo progression
C	Sq	Pembro	SD	Val	-	-	67%	-	-	"Progressive SD"
D	Ad	Pembro	SD	Val	-	-	81%	-	-	Small amount of pleural fluid
E	Ad	Nivo	SD	Val	86%	-	170%	90%	-	Nephrodrain, relatively normal kidney function (GFR 43)

Table S5.B - False positives in the remaining markers (CA125 and SCC)

ID	Diagnosis	Treatment	Response	Cohort	CA125	CEA	Cyfra	NSE	SCC	Explanation
F	Ad	Nivo	SD	Val	151%	-	-	-	-	"Progressive SD"
G	Sq	Nivo	SD	Val	51%	-	-	-	-	None
H	Ad	Nivo	SD	Train	59%	-	-	-	-	None, although suffering from brain infarction
I	Ad	Nivo	SD	Val	111%	-	-	-	-	None
J	Ad	Nivo	SD	Val	70%	-	-	-	-	Skin rash grade 2
K	Ad	Pembro	PR	Val	100%	-	-	-	-	Pseudo progression
L	Ad	Nivo	SD	Val	-	-	-	-	63%	None
M	Sq	Nivo	SD	Train	-	-	-	-	92%	Progressive pleural metastasis.

Table S5 – False positives.

A. The false positive results in the final test for Cyfra/CEA/NSE. For all the false-positive tests in the final test at 6 weeks after start, which was defined as a responder classified by non-responder with our test, with a possible explanation. B. The false positive results for the remaining markers. ID: Identification number of a patient; Response: the response after six months according RECIST criteria, with SD: Stable disease and PR: partial response ; PFS: Progression free survival in days; Ad: Adenocarcinoma; Sq: Squamous Cell Carcinoma; Nivo: nivolumab; Pembro: Pembrolizumab; Val: Validation cohort; Train: Training cohort.

Table S6 – Sub analysis SCC

Minimum value SCC		Specificity	Sensitivity	Positive predicted value
0.0 µg/L	All pathology	76.4% (67.0-83.9%)	22.9% (16.9-30.2%)	60.3% (47.2-72.2%)
	Squamous	80.8% (60.0-92.7%)	41.0% (26.0-57.8%)	76.2% (52.5-90.9%)
2.0 µg/L	All pathology	95.3% (88.8-98.3%)	9.6% (5.8-15.4%)	76.2% (52.4-90.9%)
	Squamous	88.5% (68.7-97.0%)	23.1% (11.7-39.7%)	75.0% (42.8-93.3%)
3.5 µg/L	All pathology	97.2% (91.3-99.3%)	6% (3-11.1%)	76.9% (46.0-93.8%)
	Squamous	96.2% (78.4-99.8%)	15.4% (6.4-31.2%)	85.7% (42.0-99.2%)

Sub analysis of the full cohort versus squamous cell only. µg/L: microgram per liter.

Table S7 – Patient characteristics pembrolizumab first line

Pembrolizumab first line	Non-responders (PD)	Responders (PR & SD)	
	N=8	N=23	*P-value*
Patient			
Male sex – no. (%)	4	7	*0.319*
Age (years) – mean (SD)	57.1 (SD: 6.9)	65.5 (SD: 9.4)	*0.029*
Smoking (never) – no. (%)	0	0	-
Pack years – mean (SD)	35.4 (SD 27.9)	27.6 (SD: 13.2)	*0.279*
WHO ≥ 2– no.(%)	3	1	*0.018*
Tumor characteristics			
Adenocarcinoma	6	17	*0.646*
Squamous	0	2	*0.155*
Other	2	4	-
KRAS positive	3	16	*0.115*
PD-L1 >50%	8	23	
Brain Metastasis – no.(%)	2	1	

Abbreviations: N: Number; SD: Standard Deviation; no.: Number of patients, WHO: performance-status score: World Health Organization performance status score, this is a score ranging from 0 to 5, where 0 indicates no symptom, 1 indicates mild symptoms and above 1 indicates greater disability; KRAS: Kirsten rat sarcoma viral oncogene; PD-L1: Programmed death ligand 1.

Table S8 – Results of tumor marker test at 6 weeks (5-7 weeks) for Pembrolizumab monotherapy.

Setting	Sensitivity (95%-CI)	Specificity (95%-CI)	Positive Predicted Value (95%-CI)
CEA 6 µg/L OR Cyfra 4 µg/L	25,0% (4.5-64.4%)	95,6% (76.0-99.8%)	66,7% (12.5-98.2%)
CEA 6 µg/L OR Cyfra 4 µg/L OR NSE 20 µg/L	25,0% (4.5-64.4%)	95,6% (76.0-99.8%)	66,7% (12.5-98.2%)
Cyfra 10 µg/L OR CEA 10 µg/L	20% (3-56%)	96% (80-100%)	66% (15-46%)
Cyfra 10 µg/L OR CEA 10 µg/L OR NSE 20 µg/L	20% (3-56%)	96% (80-100%)	66% (15-46%)

A total of 62 patients were treated with pembrolizumab first line. From 27 patients there was no test: 4 patients were non-evaluable, 7 patients were lost to follow up (treatment continuation in another hospital), 20 patients missed either the baseline or follow-up measurement. Data from the remaining 31 patients was used for the analysis. The same criteria were used as in the manuscript. Patient characteristics of this cohort can be found in table S.