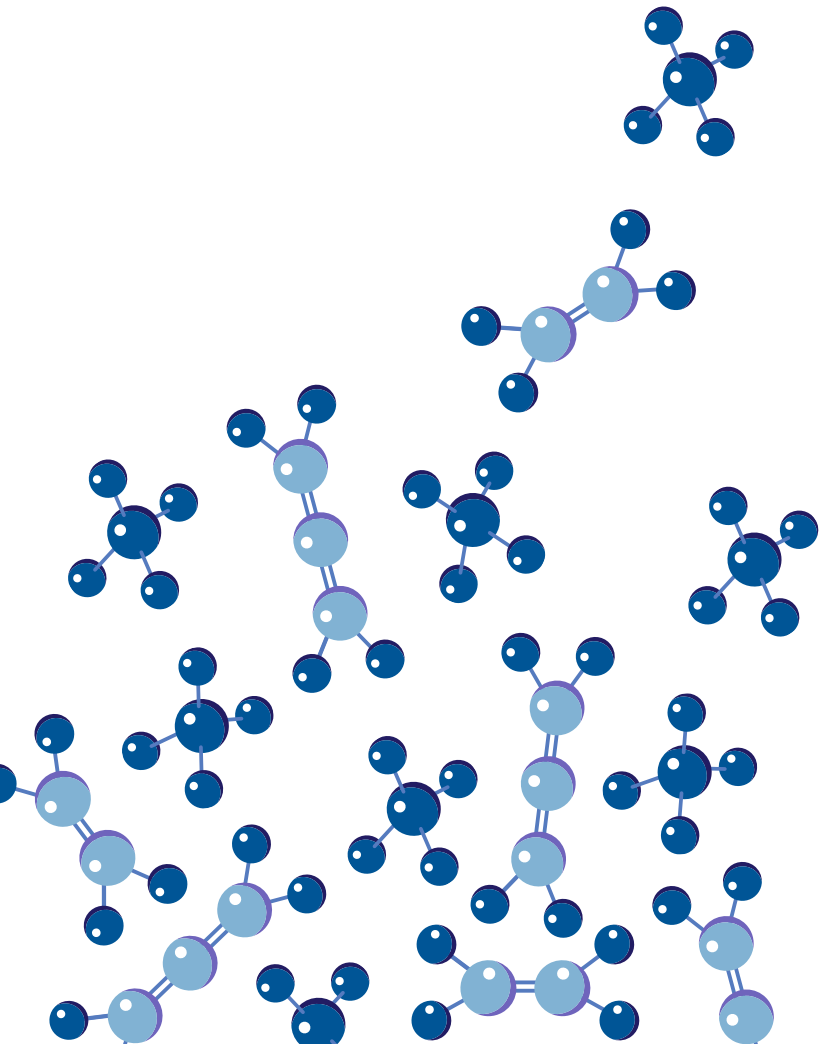
Biomarkers for the response to immunotherapy in patients with non-small cell lung cancer

Muller, M.



6 *eNose analysis for early immunotherapy response monitoring in non-small cell lung cancer*

A. Buma*, M. Muller*, R de Vries, P.J. Sterk, V. van der Noort, M. Wolf-Lansdorf, N. Farzan, P. Baas, M. van den Heuvel

Lung Cancer. 2021;160:36-43.

*Contributed equally

Abstract

Objectives: Exhaled breath analysis by electronic nose (eNose) has shown to be a potential predictive biomarker before start of anti-PD-1 therapy in patients with non-small cell lung carcinoma (NSCLC). We hypothesized that the eNose could also be used as an early monitoring tool to identify responders more accurately at early stage of treatment when compared to baseline. In this proof-of-concept study we aimed to definitely discriminate responders from non-responders after six weeks of treatment.
Materials and Methods: This was a prospective observational study in patients with advanced NSCLC eligible for anti-PD-1 treatment. The efficacy of treatment was assessed by the Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1 at 3-month follow-up. We analyzed SpiroNose exhaled breath data of 94 patients (training cohort n=62, validation cohort n=32). Data analysis involved signal processing and statistics based on Independent SamplesT-tests and Linear Discriminant Analysis (LDA) followed by Receiver Operating Characteristic (ROC) analysis.
Results: In the training cohort, a specificity of 73% was obtained at a 100% sensitivity level to identify objective responders. The Area under the Curve (AUC) was 0.95 (CI: 0.89-1.00). In the validation cohort, these results were confirmed with an AUC of 0.97 (CI: 0.91-1.00).
Conclusion: Exhaled breath analysis by eNose early during treatment allows for a highly accurate, non-invasive and low-cost identification of advanced NSCLC patients who benefit from anti-PD-1 therapy.

Keywords: non-small cell lung cancer, immunotherapy, exhaled breath analysis, non-invasive biomarker.

Introduction

The recent introduction of immune checkpoint inhibitors (ICIs) in daily clinical practice has significantly improved the 5-year survival rate in patients with metastatic non-small cell lung cancer (NSCLC) [248]. Nevertheless, results have shown that only a minority of patients experiences a relevant clinical benefit [249]. Treatment continuation is currently based on tumor dynamics evaluated by radiological imaging. However, tumor dynamics can be difficult to interpret when tumor regression occurs slowly, there is no measurable disease, or tumors even transiently progress due to inflammation [250, 251]. Since the only validated predictive biomarker tumor PD-L1 expression is fairly inaccurate, other, and preferably non-invasive, predictive biomarkers are being investigated to avoid losing valuable time and undesirable immune-related adverse events (IRAEs), and to reduce unnecessary costs [249, 252-256].

Recent studies have been exploring the use of exhaled breath analysis with "electronic noses" (eNose), which recognize gas mixtures from volatile organic compounds (VOCs). VOCs are defined as chemical compounds that have a high vapor pressure at room temperature and are a result of metabolic changes in the body [229, 257]. ENoses have been designed for classification of VOCs by pattern recognition, which can be used for probabilistic assessment of disease states. Promising results have been observed in different diseases, particularly in the field of respiratory medicine [258, 259]. Recently, *De Vries et al.* showed that exhaled breath analysis by eNose can be used before start of treatment to identify NSCLC patients that show progressive disease (PD) to anti-PD-1 therapy with 100% specificity. This way, ineffective treatment could potentially be avoided in a quarter of the patients without withholding it to those who may benefit [260]. However, still a relevant proportion of patients will ultimately not benefit. An early monitoring tool for response during treatment would be helpful to identify those patients that are more likely to benefit from alternative options.

We hypothesized that exhaled breath patterns arising from metabolic/biochemical changes induced by effective anti-PD-1 therapy in patients with NSCLC can be used to discriminate true responders from non-responders more accurately at early stage of treatment when compared to baseline. According to this hypothesis, we expect that patients with a partial response (PR) will show greater metabolic/biochemical changes compared to patients with stable disease (SD) or PD, therefore resulting in a high predictive value in identifying true response to anti-PD-1 therapy when classifying patients with PR as responders and patients with SD or PD as non-responders. This proof-of-concept study therefore aims to determine the predictive value of exhaled breath analysis by eNose for the identification of advanced NSCLC patients with PR to anti-PD-1 therapy with 100% sensitivity after six weeks of treatment.

Methods

Study population

This was a prospective observational study in adult patients with advanced NSCLC eligible for treatment with anti-PD-1 therapy. Our cohort consists of two subsets of patients: 1) Patients included in the cohort of *De Vries et al.*, who also had received a second SpiroNose measurement after six weeks of treatment (n=64), and 2) patients recruited after publication who were only treated with pembrolizumab (n=30) [260]. Patients were recruited from the thoracic oncology outpatient clinic at the Netherlands Cancer Institute (NKI) in Amsterdam and the Radboudumc Hospital in Nijmegen between August 2015 and June 2019. The patients were only included if they had received SpiroNose measurements both at baseline (defined as 0-6 weeks prior to treatment start) as well as after six weeks of treatment (defined as 4-8 weeks of treatment), and received treatment in accordance with recent literature and local guidelines [183]. Details about the "full eligibility criteria for treatment with immunotherapy in NSCLC patients" have been described by *De Vries et al. [261]*. Patients received Nivolumab or Pembrolizumab treatment every two or three weeks, respectively (Figure 1). Patients were excluded from the study if they had consumed alcohol 12 hours before the measurement, or when they were not willing to participate. Additional restrictions in eating, drinking, smoking and medication were not requested in order to make exhaled breath measurements applicable in daily clinical practice [65].

The study was approved by the ethics review board of the NKI. Details are described in the *Online Supplement*. Patients participating in the Thoracic Oncology Biobank provided written informed consent according to the Thoracic Oncology Biobank study protocol. Measurements were not performed in patients with severe shortness of breath, inability to perform a vital capacity maneuver, or inability to hold breath for five seconds. Patients were only included if they had received SpiroNose measurements both at baseline (defined as 0-6 weeks prior to treatment start) as well as after six weeks of treatment (defined as 4-8 weeks of treatment). The choice for these cut-off periods was based on our aim to include as many patients as possible.

The ethics review board of the Netherlands Cancer Institute (NKI) concluded in writing that Dutch legislation on human participation in research was not considered to be applicable, given the non-invasive nature of this study that merely added exhaled breath analysis to standard diagnostic procedures. Other clinical data (e.g. CT scan, blood tests and lung function) used in this study were collected for routine clinical practice and were subsequently handled by complying with the Dutch Personal Data Protection Act (WBP). Despite the waiver that was provided by the ethics review board, the purpose of adding the eNose to routine diagnostics was explained to the patients who all gave their oral consent.

Measurements

Within two weeks before start of treatment, blood tests and spirometry were done for toxicity monitoring, and repeated every 2 and 6 weeks respectively. For response monitoring a computed tomography (CT) scan was done within 2 weeks before start of treatment, 6 and 12 weeks after start of treatment, and repeated every 3 months. Based on the Response Evaluation Criteria in Solid Tumors (RECIST) version 1.1 criteria, tumor

dynamics evaluated in all patients with CT-imaging at baseline and at 3-month follow-up were classified as partial response (PR), stable disease (SD) or progressive disease (PD) [105]. Patients classified as PR were categorized as objective responders, while patients showing SD or PD were categorized as non-responders.

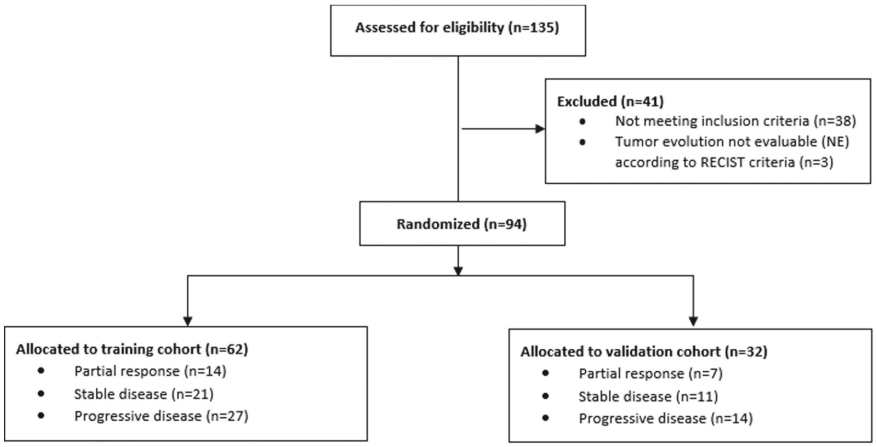


Figure 1 - CONSORT flow diagram of participants through the study.
The 38 participants that did not meet the inclusion criteria had not received SpiroNose measurements both at baseline (defined as 0-6 weeks prior to treatment start), as well as after six weeks of treatment (defined as 4-8 weeks of treatment).
Abbreviations: RECIST, Response Evaluation Criteria in Solid Tumors.

Study design

After response evaluation had been obtained for all patients, patients were randomized between a training and a validation cohort in a 2:1 ratio. Our aim was to keep both cohorts as representative as possible. Therefore, randomization was stratified according to the before mentioned response criteria at 3-month follow-up to keep an equal distribution in responses in both cohorts. Investigators were blinded to the exhaled breath data until after randomization.

In the training cohort two models for predicting response based on exhaled breath data were fitted: one using only baseline measurements (the “baseline model”) and one using both measurements collected at baseline and measurements collected after six weeks of treatment (the “on treatment model”). Then the performance of both models was evaluated in a cohort of patients not involved in the fitting of the models: the validation cohort.

Exhaled breath analysis

Exhaled breath measurements were performed using a cloud-connected eNose; SpiroNose® (Breathomix, Leiden, The Netherlands). The measurements took place the same day as the spirometry tests. The SpiroNose contains seven different cross-reactive metal oxide semiconductor (MOS) sensors and each sensor is present in duplicate both on the inside and outside of the SpiroNose. A detailed description of the SpiroNose measurement technology and breath sampling methods has been provided by *De Vries et al.* [234, 260]. The inner sensors measure the complete mixture of VOCs in exhaled breath and the outer sensors measure the ambient VOCs for background correction. Each sensor is used to determine two variables; 1) the highest sensor peak normalized to sensor 2, which is the most stable sensor, and 2) the ratio between the sensor peak and breath hold (BH) point [239, 261]. Measurements were performed in duplicate with a 2-minutes interval at baseline and after six weeks of treatment.

Data processing

Processing of the eNose sensor signals included filtering, de-trending, ambient correction and peak detection by the standard eNose software as described by *De Vries et al*. [65, 234, 260]. A .csv file was used to store the selected parameters (sensor peak- and peak/BH ratios) resulting from the signal processing and served as the source document for statistical analysis.

Statistical analysis

Patient and tumor characteristics

Data-analysis was performed using MatLab (Version 2019b) and IBM SPSS Statistics (Version 26) and is explained in the *Online Supplement*.

Patient and tumor characteristics were described and compared between responders and non-responders, for both cohorts separately, considering a p-value <0.05 as statistically significant. Continuous variables were reported as means (SD) or medians (IQR) for normally and non-normally distributed data, respectively. Categorical variables were reported as ratios. Intergroup comparisons were performed using One-way ANOVA tests, Kruskal Wallis tests or Chi-squared tests.

Sample size calculation

Due to logistic reasons the total number of patients in the cohort was fixed. However, a calculation for the training and validation cohort was possible. Our aim was to make the training cohort as large as possible, while still having sufficient patients in the validation cohort to draw meaningful conclusions. We decided on forehand that a model developed in the training cohort would be considered successful if the two-sided 95% DeLong confidence interval around the AUC as established in the validation cohort would be entirely above 0.70, thus clinically relevantly far removed from the null-value of 0.5. Furthermore, our aim was to develop a biomarker in the training cohort that would be as accurate as the biomarker of *De Vries et al*. (which had an AUC of 0.85), to show the added value of "on treatment" breath profiles [260]. With this in mind, we decided to randomize the patients in a 2:1 ratio (training cohort n=62, validation cohort n=32) if simulations would show that 32 patients in the validation cohort would still yield sufficient accuracy. In order to determine (prior to randomization) whether a validation cohort of 32 patients

would yield sufficient accuracy to declare a marker as accurate as the one developed by *De Vries et al*. "successful" according to the above criterion, we randomly drew 10.000 virtual validation cohorts of 32 patient each, and computed the AUCs with confidence intervals of the actual *De Vries et al.* biomarker in each of these cohorts. The mean lower bound of these confidence intervals was 0.766, indicating that a 2:1 randomization would indeed yield a sufficiently large validation cohort according to our pre-specified criterion.

Exhaled breath analysis: training cohort
Since the model of *De Vries et al.* aims to identify patients classified as PD with 100% specificity, our training cohort was used to make a new predictive model based on baseline measurements only (the "baseline model") to identify patients classified as PR with 100% sensitivity. Furthermore, a predictive model was composed that included measurements performed both at baseline and after six weeks of treatment (the "on treatment model") to be able to determine the additional value of measurements performed early during treatment. Independent Samples T-tests and Linear Discriminant Analysis (LDA) were used to identify sensor values with the highest contribution to the discrimination of patients classified as PR and patients classified as SD or PD. For the baseline model, only baseline sensor values were included in both analyses. For the on treatment model, baseline sensor values and sensor values obtained after six weeks of treatment were included. The suffixes _6, _absdif and _reldif are used to indicate the sensor variables "value after six weeks of treatment", "absolute difference" and "relative difference", respectively. Details regarding the construction of the models are provided in the *Online Supplement*. Receiver Operating Characteristic (ROC) curves were constructed for the composed predictive models, and associated Area Under the Curves (AUCs) and specificities when focusing on a 100% sensitivity to identify patients classified as objective responders were calculated.

In the training cohort, two predictive models were composed to identify at baseline and after six weeks of treatment patients with an objective response to anti-PD-1 therapy with 100% sensitivity, respectively. Both models were examined in the validation cohort to test external validity. Variables that were considered for inclusion in the models were 1) all normalized sensor peaks and peak/ breath hold (BH) ratios at baseline, 2) all normalized sensor peaks and peak/BH ratios after six weeks of treatment, and 3) all absolute and relative differences between baseline and six weeks of treatment in sensor peaks and peak/BH ratios. Absolute sensor differences were calculated by subtracting sensor values measured after six weeks of treatment from sensor values measured at baseline for each sensor. Relative sensor differences were calculated by dividing the calculated absolute sensor differences by the sensor values measured at baseline for each sensor. The characters _6, _absdif and _reldif are used to indicate the sensor variables "value after six weeks of treatment", "absolute difference" and "relative difference", respectively.

Firstly, the values of all predictors were compared univariately between patients with a partial response (PR) and patients with stable disease (SD) or progressive disease (PD) by means of an Independent Samples T-test. Secondly, various multivariate models created by linear discriminant analysis (LDA) were considered. We decided to include

absolute over relative sensor differences in het multivariate models, since relative sensor differences are based on ratios and therefore more prone to potential errors. Since for any sensor, any two of the variables "value at baseline", "value after six weeks of treatment" and "absolute difference" are sufficient to compute the remaining variable, we barred these variables from entering the models with more than two at a time. Variable selection for the final LDA models was based on the performance of the predictors in the univariate and multivariate models in combination with our aim of including a maximum of six predictors into each model to reduce the risk of overfitting.

Exhaled breath analysis: validation cohort
For each patient in the validation cohort the values of the discriminant functions given by the baseline model and the on treatment model (composed in the training cohort) were calculated. These values were used to construct ROC curves in the validation cohort. To test external validity, the AUCs of these ROC curves were compared to the AUCs of the ROC curves composed in the training cohort and to the fixed boundary of 0.7. The predictive accuracy of both models was established in the validation cohort based on AUC and on specificity when focusing on a 100% sensitivity to identify patients with an objective response.

Finally, the discriminant scores calculated from each model for each patient were converted into prediction scores to facilitate interpretation in daily clinical practice. A cut-off point was selected in the training cohort to identify objective responders with 100% sensitivity after six weeks of treatment. This cut-off point was translated into a prediction score and examined in the validation cohort by calculating the associated sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). Survival analyses with Kaplan Meier curves were performed to assess the relation between our results and survival.

Results

In total, 94 advanced NSCLC patients were enrolled in this study (Figure 1), from who 64 were included in the study of *De Vries et al*. [260]. They were randomly assigned to the training (n=62) or validation cohort (n=32) (Table 1), according to the before mentioned criteria, resulting in 62 patients in the training cohort and 32 patients in the validation cohort. All baseline measurements were performed with a median of 2 weeks (range: 0-6 weeks) before treatment. The follow-up measurements were performed with a median of 6 weeks (range: 4-8 weeks) and 12 weeks (range: 10-14 weeks) for the eNose and CT-scan respectively. In the validation cohort, a significant difference was seen in choice of treatment between the three groups (p=0.01). Patients showing SD or PD after three months of treatment were more often treated with Nivolumab compared to Pembrolizumab. No significant differences were seen in any other baseline characteristic between the three groups.

Table 1 - Baseline characteristics of the training and validation cohort.

	Training (n=62)			Validation (n=32)		
	PR (n=14)	**SD (n=21)**	**PD (n=27)**	**PR (n=7)**	**SD (n=11)**	**PD (n=14)**
Patient						
Age in years, mean (SD)	64.7 (7.4)	66.7 (7.9)	65.63 (9.3)	64.3 (9.4)	65.2 (7.2)	59.4 (9.8)
Gender (males), N (%)	6 (42.9)	13 (61.9)	14 (51.9)	3 (42.9)	6 (54.5)	8 (57.1)
BMI, median (IQR)	22.8 (20.7-27.2)	25.3 (23.7-30.4)	27.0 (23.2-31.4)	24.9 (22.7-28.3)	26.6 (23.9-31.9)	25.0 (20.4-26.8)
FEV1%, median[a] (IQR)	2.0 (1.6-2.3)	1.6 (1.4-2.3)	1.9 (1.5-2.6)	1.8 (1.7-2.0)	1.9 (1.6-2.0)	1.9 (1.6-2.4)
WHO performance[b] (≥2), N (%)	1 (7.7)	0 (0.0)	2 (7.4)	0 (0.0)	2 (20.0)	1 (7.1)
Ethnicity (Caucasian)[b], N (%)	13 (100.0)	18 (94.7)	25 (92.6)	7 (100.0)	9 (90.0)	14 (100.0)
Smoking status, N (%)						
Never smoker	0 (0.0)	2 (9.5)	2 (7.4)	1 (14.3)	2 (18.2)	2 (14.3)
Current smoker	3 (21.4)	6 (28.6)	6 (22.2)	2 (28.6)	2 (18.2)	5 (35.7)
Ex-smoker	11 (78.6)	13 (61.9)	19 (70.4)	4 (57.1)	7 (63.6)	7 (50.0)
Pack-years[c], median (IQR)	25.0 (15.0-40.0)	31.0 (7.0-45.0)	30.0 (20.0-50.0)	41.0 (32.5-52.0)	27.0 (10.5-32.0)	29.0 (15.0-40.0)
Tumor characteristics						
Histologyb, N (%)						
AC	9 (69.2)		19 (70.4)	3 (42.9)	6 (54.5)	8 (57.1)
SCC	1 (7.7)	4 (22.2)	6 (22.2)	2 (28.6)	4 (36.4)	2 (14.3)
Other	3 (23.1)	1 (5.6)	2 (7.4)	2 (28.6)	1 (9.1)	4 (28.5)
Mutation statusb, N (%)						
KRAS positive	6 (46.2)	9 (47.4)	11 (50.0)	1 (20.0)	3 (33.3)	4 (33.3)
EGFR positive	0 (0.0)	0 (0.0)	1 (4.5)	2 (40.0)	0 (0.0)	1 (9.1)
BRAF positive	1 (10.0)	0 (0.0)	1 (4.5)	0 (0.0)	0 (0.0)	0 (0.0)

Table 1 - Baseline characteristics of the training and validation cohort. *(Continued)*

	Training (n=62)			Validation (n=32)		
	PR (n=14)	**SD (n=21)**	**PD (n=27)**	**PR (n=7)**	**SD (n=11)**	**PD (n=14)**
PD-L1 expressiond, N (%)						
Negative (0%)	1 (7.1)	9 (42.9)	6 (22.2)	0 (0.0)	3 (27.3)	6 (42.9)
Weak positive (<50%)	0 (0.0)	1 (4.8)	4 (14.8)	0 (0.0)	1 (9.1)	2 (14.3)
Strong positive (>50%)	8 (57.1)	8 (38.1)	4 (14.8)	4 (57.1)	2 (18.2)	0 (0.0)
Unknown	5 (35.7)	3 (14.3)	13 (48.1)	3 (42.9)	5 (45.5)	6 (42.9)
Cancer stage III, N (%)	1 (7.1)	3 (14.3)	4 (14.8)	1 (14.3)	3 (27.3)	0 (0.0)
Treatment						
Current[e], N (%)						
Nivolumab	9 (64.3)	13 (61.9)	19 (70.4)	3 (42.9)	9 (81.8)	13 (92.9)
Pembrolizumab	5 (35.7)	8 (38.1)	8 (29.6)	4 (57.1)	2 (18.2)	1 (7.1)
Line of treatment, N(%)						
1st line	5 (35.7)	6 (28.6)	3 (11.1)	2 (28.6)	3 (27.3)	1 (7.1)
≥ 2 line	9 (64.3)	15 (71.4)	24 (88.9)	5 (71.4)	8 (72.7)	13 (92.9)

Abbreviations: BMI, body mass index; FEV1, forced expiratory volume in 1 second; WHO, world health organization; COPD, chronic obstructive pulmonary disease; GOLD: Global Initiative for Chronic Obstructive Lung Disease; AC, adenocarcinoma; SCC, squamous cell carcinoma; EGFR, epidermal growth factor receptor.

[a] One patient (SD) missing FEV1% at baseline in both training and validation cohort.

[b] Not available for all patients at baseline in both training and validation cohort, percentage shown in percentage of known cases.

[c] Six patients (PR n=1, SD n=3, PD n=2) missing pack-years in training cohort.

[d] Significant difference between PR, SD and PD in both training (p=0.03) and validation cohort (p=0.002).

[e] Significant difference between PR, SD and PD in validation cohort (p=0.01).

Exhaled breath analysis: training cohort

Sensor 3 (p=0.001), sensor 5_BH (p<0.001), sensor 3_6 (p<0.001), sensor 4_6 (p=0.05), sensor 2_BH_6 (p=0.04), sensor 5_BH_6 (p=0.005), sensor 3_reldif (p<0.001) and sensor 3_absdif (p<0.001) significantly differed between patients classified as PR and patients classified as PD or SD, which is shown in *Supplementary Figure S1* and *Supplementary Figure S2* for sensor 3_absdif. Results obtained from LDA are described in the *Online Supplement*.

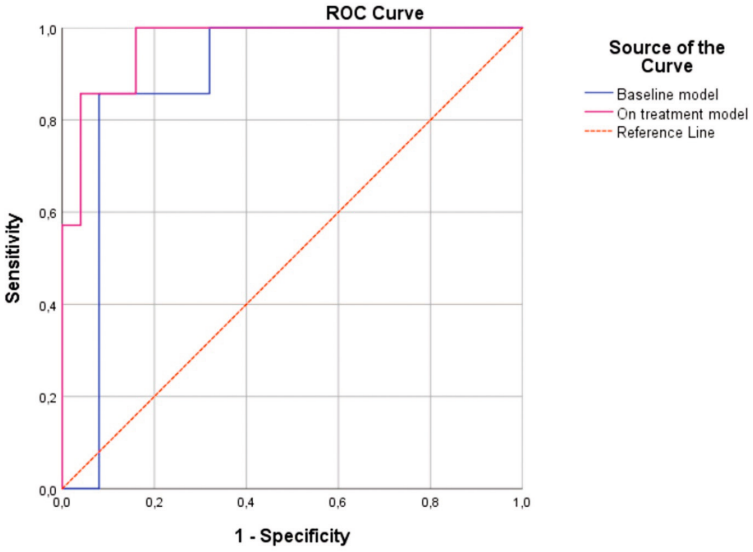


Figure 2 - ROC curves composed for the baseline model and model 2 predictive for the identification of patients showing an objective response to anti-PD-1 therapy in the validation cohort.
Baseline model: sensor 3, sensor 3_BH, sensor 5_BH and sensor 6_BH.
Model 2: sensor 4, sensor 6, sensor 1_6, sensor 6_6 and sensor 3_absdif.
Abbreviations: ROC, receiving operating characteristic; BH, breath hold; sensor 1_6, sensor value measured by sensor 1 after six weeks of treatment; sensor 6_6, sensor value measured by sensor 6 after six weeks of treatment; S3_absdif, sensor value difference between six weeks of treatment and baseline measured by sensor 3.

The first model (baseline model), based on baseline measurements only, reached a specificity of 54% when requiring 100% sensitivity and had a ROC-AUC of 0.81 (CI: 0.71-0.92) (Figure 2). In the second model (on treatment model), that included measurements performed both at baseline and after six weeks of treatment, a specificity of 73% at 100% sensitivity and a ROC-AUC of 0.95 (CI: 0.89-1.00) was reached (Figure 2). Details on the composition of the two models are provided in the *Online Supplement.*

Exhaled breath analysis: validation cohort

In the validation cohort, the baseline model reached a specificity of 68% when requiring 100% sensitivity and a ROC-AUC of 0.89 (CI: 0.76-1.00). The on treatment model reached a specificity of 84% at 100% sensitivity and a ROC-AUC of 0.97 (0.91-1.00). The baseline model reached a specificity of 68% at a ROC-AUC of 0.89 (CI: 0.76-1.00) (Figure 2 and *Supplementary Table S1*).

The equation below resulted from LDA on the on treatment model and can be used for response prediction in future patients. Details on the baseline model and the mathematical derivation of the models are provided in the *Online Supplement*.

Prediction score (patient) on treatment model =

$$\frac{1}{1 + e^{3.1180 + 4.6260 * S4 - 4.5276 * S6 + 1.0906 * S1\_6 - 0.9524 * S6\_6 + 26.0143 * S3\_absdif}}$$

The above equations were used to convert discriminant scores into prediction scores in the validation cohort (Figure 3). The cut-off point selected in the training cohort, aiming not to withhold anti-PD-1 therapy to objective responders after six weeks of treatment, corresponded to a prediction score of ≥0.14 for membership to the PR group. Patients with a prediction score below this cut-off point were classified as non-responders. In the validation cohort, this cut-off point showed a sensitivity of 100%, a specificity of 76%, a PPV of 54%, and a NPV of 100% to identify objective responders. Kaplan Meier survival curves are shown in *Supplementary Figure S3* and *Supplementary Figure S4*.

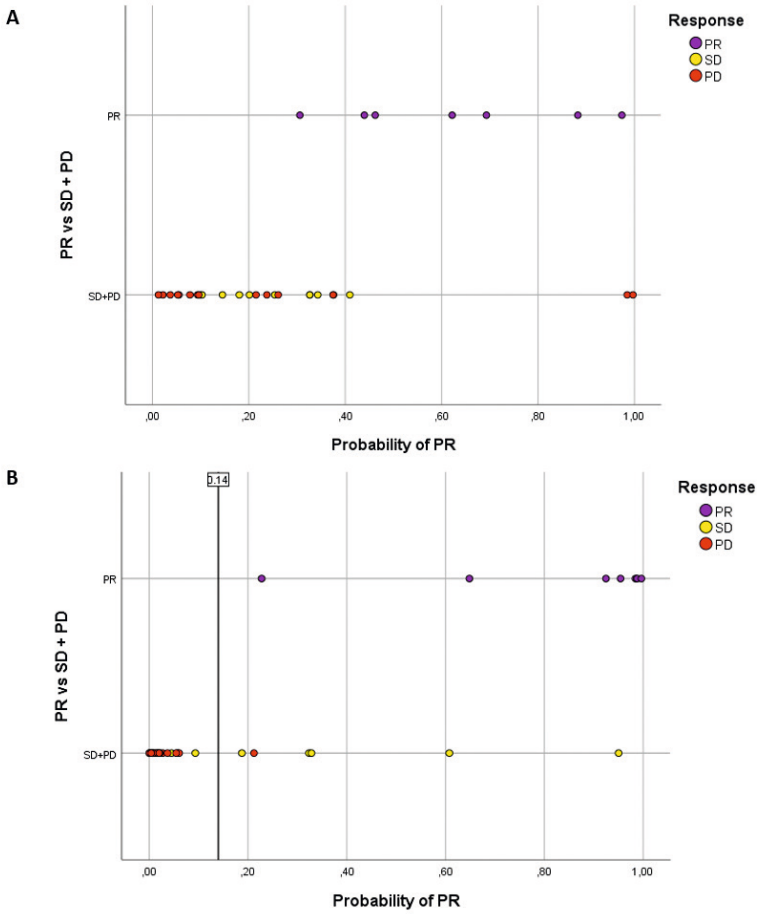


Figure 3 - Scatterplots representing the prediction scores calculated with the baseline model and model 2 for each patient (n=32) in the validation cohort.
A) Baseline model.
B) Model 2: All patients showing PR to anti-PD-1 therapy (n=7) are correctly classified when applying a cut-off point of ≥0.14 for membership to the objective response group. 5 out of 11 patients classified as SD and 1 out of 14 patients classified as PD are incorrectly classified.
Abbreviations: PR, partial response; SD, stable disease; PD, progressive disease.

Discussion

This prospective observational study shows that SpiroNose exhaled breath analysis can be used to identify advanced NSCLC patients with an objective response to anti-PD-1 therapy more accurately at early stage of treatment when compared to baseline as part of routine assessment during early treatment monitoring in daily clinical practice. Results obtained in the training cohort were confirmed in the validation cohort with an AUC of 0.97 (CI: 0.91-1.00).

To the best of our knowledge, this is the first study that has applied eNose technology to identify advanced NSCLC patients with a PR to anti-PD-1 therapy and to investigate the potential additional value of SpiroNose exhaled breath measurements early during treatment. Our study extends the work of *De Vries et al.*, who investigated whether sensitivity to anti-PD-1 therapy in patients with advanced NSCLC might be reflected by a distinct exhaled breathprint. They showed that SpiroNose exhaled breath analysis could indeed be used to discriminate at baseline patients showing PD from patients showing PR or SD, and with a superior predictive performance than obtained with the current clinical standard biomarker PD-L1.

In this study, we obtained an increased discriminative potential for the identification of patients with PR when applying the on treatment model when compared to the baseline model (*Supplementary Table S1*). After six weeks of treatment, patients classified as PR showed a distinct clustering of prediction scores towards higher probabilities of an objective response, while patients classified as PD showed a distinct clustering towards lower probabilities of an objective response when compared to baseline. Patients categorized as SD, on the other hand, showed an increased spread in prediction scores, with the majority of scores falling back to low probabilities of an objective response (Figure 3). Based on these results, one could argue that this increased discriminative potential after treatment initiation might be partly driven by VOCs that arise from metabolic/biochemical changes induced by anti-PD-1 therapy. This would imply that a direct treatment effect could be monitored through exhaled breath. We therefore suggest that the on treatment model could therefore not only be used as a predictive biomarker to identify patients exhibiting primary resistance as early as six weeks following treatment initiation, but also as a real-time monitoring tool for therapeutic efficacy during follow-up to identify patients developing secondary resistance during course of treatment. However, we suggest this model first to be validated in an external, prospective, and preferably multicenter, validation cohort to confirm the predictive value obtained in our study. Subsequently, application of the model in daily clinical practice should be investigated in combination with other current biomarkers (e.g. clinical condition, serum markers, radiological imaging, histopathology, etc.) to help therapeutic decision-making during course of treatment. We expect that this approach will allow for an earlier and more precise identification of non-responding patients during anti-PD-1 therapy when compared to current follow-up care, and subsequently help to avoid undesirable events of treatment and losing valuable time in a higher percentage of these patients.

As eNoses have been designed for probabilistic assessment of VOCs, based on pattern recognition algorithms, It remains to be determined which specific metabolic/biochemical pathways contribute to the associations between the measured VOC-patterns and the patient response evaluation. When looking at our composed predictive models and the model of *De Vries et al.*, we observe that sensor 3, which is most sensitive to methane and natural gas, consistently has a major contribution to the discriminative performance of all models [260]. The model of *De Vries et al.* aimed to identify patients showing PD with a 100% specificity, classifying patients as PR or SD as responders. Our study aimed to improve the applicability of SpiroNose exhaled breath analysis in daily clinical practice. Mean sensor values calculated for sensor 3 for each response group showed that patients classified as PR had the highest mean sensor value at baseline, while patients with PD had the lowest (*Supplementary Table S3* and *Supplementary Table S4*). Furthermore, the SD group showed a mean sensor value more similar to the mean sensor value calculated for patients with PR. After six weeks of treatment, however, exhaled breath patterns distinctly differed for each response group. Patients categorized as PR showed a significant decrease in measured sensor values and had the lowest mean sensor value, while patients with PD had the highest. Interestingly, patients classified as SD showed a mean sensor value more similar to the mean sensor value calculated for the PD group. One could therefore speculate that the majority of patients classified as SD exhibit slow tumor progression during treatment. Classifying patients with PR or SD as responders during treatment would therefore have resulted in a smaller difference in mean sensor value between the responder and non-responder group, resulting in a lower predictive value in identifying true response to anti-PD-1 therapy when applying the on treatment model. We therefore believe that the current classification of responders brings the evaluation of response closest to the actual response occurring within the patient during treatment. Applying a predictive model that aims to identify VOC pattern changes occurring in true responders could therefore be more sensitive in identifying patients within the SD group who have a delayed but potentially durable response to anti-PD-1 therapy. This way, patients could be classified as responders or non-responders instead of PR/SD/PD, facilitating the decision to continue, stop or adapt treatment during current and future immunotherapy options [262]. Furthermore, we suggest that it should be investigated which individual VOCs contribute to the discrimination between responders and non-responders in order to draw conclusions about which specific metabolic/biochemical pathways are associated with response. Insight into these molecular mechanisms could then be used to improve the SpiroNose as biomarker tool.

One could argue that pattern recognition rather than identification of VOCs is an intrinsic limitation of using eNoses. Exhaled breath analysis technology comprises multiple methods for breath sampling [229]. Different studies have been able to identify multiple compounds associated with lung cancer by using methods that aim to detect, identify, and quantify specific, individual chemical compounds in exhaled breath [263-265]. However, these methods have shown some practical disadvantages, which makes them less suitable for clinical implementation yet [65, 228, 239, 260, 266]. In the present study, we were able to accurately identify true responders to anti-PD-1 therapy by using an eNose based on cross-reactive nonspecific sensor arrays without requiring

any restrictions except from alcohol consumption12 hours before the measurement. In addition, we were able to identify true responders in a relatively heterogenous population of patients (Table 1). This could imply that the SpiroNose identifies breath patterns associated with response that are not influenced by baseline characteristics and lifestyle of patients, which increases its external validity. Since it remains unclear which intrinsic and extrinsic factors determine the breath print typically for a response to immunotherapy and which set of VOC's characterize this breath print, we believe that the use of an eNose based on pattern recognition allows for a less error-prone, and therefore more accurate, approach for identifying responding patients as part of routine assessment in daily clinical practice when compared to individual VOC detection methods.

A limitation of our study might be the response categorization based on conventional radiological response criteria. Pseudoprogression, which is defined by a transient increase in tumor burden followed by a delayed decrease in tumor size, is considered one of the unusual response patterns when assessing efficacy of immunotherapy by radiological imaging and might result in incorrect classification of a subset of responders [251, 267]. Since the incidence of pseudoprogression in NSCLC patients is thought to be less than 5%, we believe the risk of misclassification bias in our study to be extremely low [267].

Conclusions

In conclusion, results obtained in the present study show that exhaled breath analysis by eNose allows for a highly accurate, non-invasive and low-cost identification of advanced NSCLC patients with an objective response to anti-PD-1 therapy as part of a routine assessment during early treatment monitoring in daily clinical practice. The clear advantage of such an identification is that application of ineffective treatments can be avoided in a higher percentage of non-responding patients, thereby preventing undesirable events and reducing unnecessary costs. Importantly, our study also paves the way for optimizing the clinical application of eNose exhaled breath analysis in patients with advanced NSCLC.

Supplemental material

The supplemental material:

Content included in this thesis:
Supplemental results
Figure S1 - Sensor value sensor 3 between baseline and six weeks of treatment
Figure S2 - Sensor value after six weeks of treatment
Figure S3 - Kaplan Meier curve showing the overall survival (OS)
Figure S4 - Kaplan Meier curve showing the progression free survival (PFS)
Table S1 - Predictive performance of the SpiroNose in both training and validation cohort.
Table S2 - Sensor values measured by sensor 3 and prediction scores
Table S3 - Mean sensor values calculated for S3, S3_6, and S3_absdif
Table S4 - Mean sensor values and difference in mean sensor values calculated for S3, S3_6, and S3_absdif

Results

Exhaled breath analysis: training cohort
Sensor 3, sensor 4, sensor 5_BH, sensor 6_BH, sensor 1_6, sensor 6_6, sensor, sensor 3_BH_6 and sensor 6_BH_6 resulted to be the sensors with the highest discriminative potential based on LDA when including both baseline sensor values and sensor values measured after six weeks of treatment. ROC curves constructed for sensor 3_reldif and sensor 3_absdif showed a large overlap between both curves, justifying our decision to include only absolute sensor signal differences when performing LDA. The best predictive model (model 1) was composed with sensor 4, sensor 6, sensor 3_BH_6, sensor 5_BH_6, sensor 6_BH_6 and S3_absdif and showed a specificity of 77% when focusing on a 100% sensitivity to identify patients classified with an objective response at an Area Under the Curve (AUC) of 0.97 (CI: 0.92-1.00). Model 2 consisted of sensor 4, sensor 6, sensor 1_6, sensor 6_6 and sensor 3_absdif. This model reached a specificity of 73% at an AUC of 0.95 (CI: 0.89-1.00) when focusing on a sensitivity of 100% to identify objective responders. Model 3 was composed with sensor 3_BH_6, sensor 5_BH_6 sensor 6_BH_6, and sensor 3_absdif, and reached a specificity of 71% at an AUC of 0.96 (CI: 0.91-1.00) (*Supplementary Table S1*).

When including only baseline sensor values, LDA showed a high contribution of sensor 3, sensor 3_BH, and sensor 6_BH to the discrimination between patients classified as PR and patients classified as SD or PD. Since the performed Independent Samples

T-test showed that sensor 3 and sensor 5_BH significantly differed between patients showing PR and patients showing SD or PD at baseline, our predictive baseline model was composed with sensor 3, sensor 3_BH, sensor 5_BH and sensor 6_BH. This model reached a specificity of 54% at a ROC-AUC of 0.81 (CI: 0.71-0.92) when focusing on a 100% sensitivity to identify patients classified as objective responders (*Supplementary Table S1*).
Exhaled breath analysis: validation cohort
In the validation cohort, prediction scores were used to visualize the discriminative potential for the baseline model and model 2 ("on treatment model"). In the context of constructing a ROC curve, the calculated prediction scores contain exactly the same information as the discriminant scores [260].
However, prediction scores facilitate interpretation in daily clinical practice.

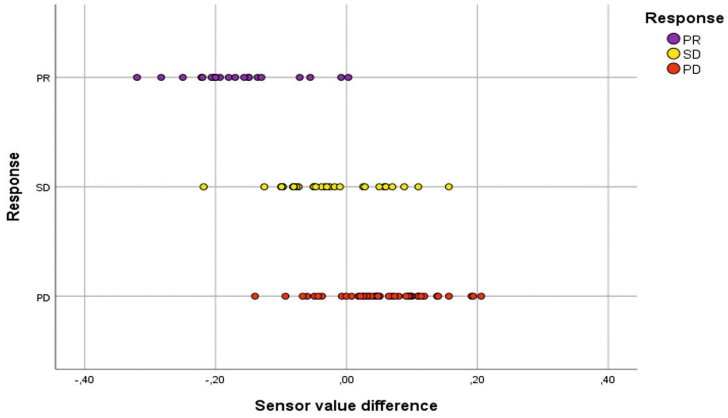


Figure S1 - Sensor value differences measured by sensor 3 between baseline and six weeks of treatment for each patient (n=94). 20 out of 21 patients classified as PR, 21 out of 32 patients classified as SD, and 8 out of 41 patients classified as PD show decreased sensor values after six weeks of treatment when compared to baseline. One patient classified as PR and one patient classified as PD show stable measured sensor values over time. All other patients show an increase in measured sensor value. Abbreviations: PR, partial response; SD, stable disease; PD, progressive disease.

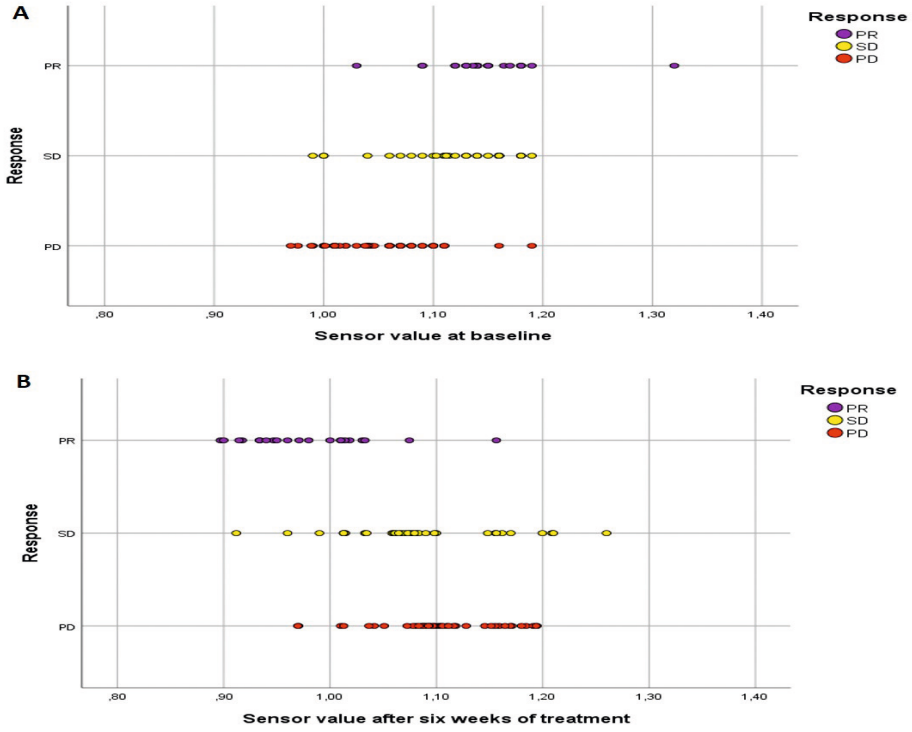
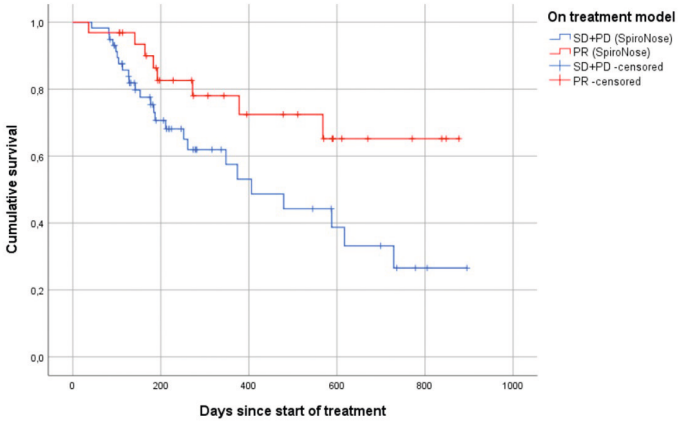


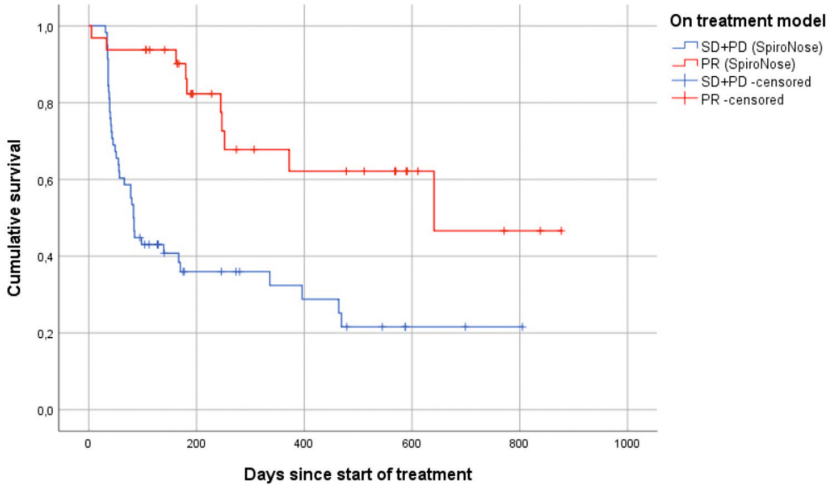
Figure S2 - Sensor values measured by sensor 3 at (A) baseline and (B) after six weeks of treatment for each patient (n=94). Abbreviations: PR, partial response; SD, stable disease; PD, progressive disease.



No. at risk	0	200	400	600	800	1000
PR	32	20	12	6	3	0
SD+PD	58	29	12	7	2	0

Figure S3 - Kaplan Meier curve showing the overall survival (OS) analysis of the two groups (PR vs. SD+PD) when applying the on treatment model (with a cut-off of 0.14) of 90 advanced NSCLC patients.

The mean overall survival was 672 days (CI: 553-791 days) for patients classified as PR and 482 days (CI: 378-585 days) for patients classified as SD or PD, according to the SpiroNose test. The survival curve showed a significant difference with a log rank (Mantel-Cox) test between objective responders and patients classified as SD or PD according to their group based on the SpiroNose (p=0.03). Abbreviations: PR, partial response; SD, stable disease; PD, progressive disease; NSCLC, non-small cell lung cancer; CI, confidence interval.



No. at risk	0	200	400	600	800	1000
PR	32	18	11	5	2	0
SD+PD	58	13	8	2	1	0

Figure S4 - Kaplan Meier curve showing the progression free survival (PFS) analysis of the two groups (PR vs. SD+PD) when applying the on treatment model (with a cut-off of 0.14) of 90 advanced NSCLC patients.

The mean progression free survival was 586 days (CI: 449-724 days) for patients classified as PR and 275 days (CI: 189-362 days) for patients classified as SD or PD, according to the SpiroNose test. The survival curve showed a significant difference with a log rank (Mantel-Cox) test between objective responders and patients classified as SD or PD according to their group based on the SpiroNose (p<0.001).

Abbreviations: PR, partial response; SD, stable disease; PD, progressive disease; NSCLC, non-small cell lung cancer; CI, confidence interval.

Table S1 - Predictive performance of the SpiroNose for the identification of patients showing an objective response to anti-PD-1 therapy in both training and validation cohort.

Model	N	Cut-off point	Sensitivity (%)	Specificity (%)	AUC (95% CI)
Training					
Baseline model	14 vs. 48	0.1080	100	54	0.81 (0.71-0.92)
On treatment model	14 vs. 48	-0.2770	100	73	0.95 (0.89-1.00)
Validation					
Baseline model	7 vs. 25	-0.5794	100	68	0.89 (0.76-1.00)
On treatment model	7 vs. 25	-0.5575	100	84	0.97 (0.91-1.00)

Abbreviations: N, number; AUC, area under the curve; CI, confidence interval.

Table S2 - Sensor values measured by sensor 3 and prediction scores calculated with the baseline model and on treatment model for each patient (PR n=7, SD n=11, PD n = 14) in the validation cohort.

Patient ID	Response	S3	S3_6	S3_ absdif	P score baseline model	P score on treatment model	Difference[a]	OS[b]	PFS[b]
1	PR	1.18	0.90	-0.28	0.68	0.98	+0.30	478	478
2	PR	1.15	0.90	-0.25	0.88	0.99	+0.11	106	106
3	PR	1.14	1.01	-0.13	0.46	0.64	+0.18	189	189
4	PR	1.32	1.00	-0.32	0.97	1.00	+0.03	838	838
5	PR	1.14	0.91	-0.23	0.61	0.92	+0.31	771	771
6	PR	1.17	0.95	-0.22	0.42	0.95	+0.53	228	228
7	PR	1.13	1.07	-0.06	0.30	0.22	-0.08	591	591
8	SD	1.13	0.91	-0.22	0.36	0.95	+0.59	395	372
9	SD	1.00	1.16	+0.16	0.11	0.00	-0.11	895	469
10	SD	1.11	1.01	-0.10	0.25	0.32	+0.07	511	511
11	SD	1.08	1.07	-0.01	0.18	0.09	-0.09	588	588
12	SD	1.10	1.06	-0.04	0.14	0.18	+0.04	197	180
13	SD	1.12	1.07	-0.05	0.33	0.04	-0.29	479	479
14	SD	1.16	1.03	-0.13	0.31	0.60	+0.29	343	245
15	SD	1.16	1.08	-0.08	0.38	0.32	-0.06	183	182
16	SD	1.16	1.21	+0.05	0.31	0.02	-0.29	219	139
17	SD	1.07	1.10	+0.03	0.20	0.02	-0.18	778	396
18	SD	0.99	0.96	-0.03	0.11	0.06	-0.05	-	-
19	PD	1.04	1.10	+0.06	0.09	0.01	-0.08	406	36
20	PD	1.07	1.08	+0.01	0.26	0.03	-0.23	374	167
21	PD	1.00	1.19	+0.19	0.04	0.00	-0.04	112	34
22	PD	1.16	1.09	-0.07	0.36	0.21	-0.15	36	33
23	PD	1.11	1.07	-0.04	0.23	0.06	-0.17	132	39
24	PD	1.08	1.13	+0.05	0.99	0.01	-0.98	282	55
25	PD	1.08	1.11	+0.03	1.00	0.02	-0.98	211	57
26	PD	1.06	1.16	+0.10	0.05	0.01	-0.04	112	112
27	PD	1.08	1.04	-0.04	0.02	0.04	0.02	252	31
28	PD	1.09	1.12	+0.03	0.07	0.05	-0.02	127	36
29	PD	1.01	1.08	+0.07	0.01	0.00	-0.01	113	39
30	PD	1.00	1.09	+0.09	0.10	0.01	-0.09	206	43
31	PD	1.09	1.11	+0.02	0.21	0.02	-0.19	729	35
32	PD	1.04	1.15	+0.11	0.05	0.00	-0.05	187	51

Abbreviations: PR, partial response; SD, stable disease; PD, progressive disease; S3, sensor value measured by sensor 3 at baseline; S3_6, sensor value measured by sensor 3 after six weeks of treatment; S3_absdif, absolute sensor value difference between sensor values measured by sensor 3 at baseline and after six weeks of treatment; P, prediction; OS, overall survival; PFS, progression free survival.

[a] *Difference between prediction scores calculated with the baseline model and on treatment model.*

[b] *Time unit is days.*

Table S3 - Mean sensor values calculated for S3, S3_6, and S3_absdif for each response group.

Advanced NSCLC (N=94)	
PR	(n=21_
S3	1.15
S3_6	0.98
S3_Absdiff	-0.17
SD	(n=32)
S3	1.11
S3_6	1.09
S3_Absdiff	-0.03
PD	(n=41)
S3	1.05
S3_6	1.10
S3_Absdiff	+0.05

Abbreviations: S3, sensor value measured by sensor 3 at baseline; S3_6, sensor value measured by sensor 3 after six weeks of treatment; S3_absdif, absolute sensor value difference between sensor values measured by sensor 3 at baseline and after six weeks of treatment; NSCLC, non-small cell lung cancer; PR, partial response; SD, stable disease; PD, progressive disease.

Table S4 - Mean sensor values and difference in mean sensor values calculated for S3, S3_6, and S3_absdif for the group of objective responders and the group of patients classified as SD or PD.

	Advanced NSCLC (n=94)		
	PR (n=21)	**SD+PD (n=73)**	**Difference**
S3	1.15	1.08	+0.07
S3_6	0.98	1.10	-0.12
S3_absdiff	-0.17	0.02	-0.18

Abbreviations: S3, sensor value measured by sensor 3 at baseline; S3_6, sensor value measured by sensor 3 after six weeks of treatment; S3_absdif, absolute sensor value difference between sensor values measured by sensor 3 at baseline and after six weeks of treatment; NSCLC, non-small cell lung cancer; PR, partial response; SD, stable disease; PD, progressive disease.