Towards overtreatment-free immunotherapy: Using genomic scars to select treatment beneficiaries in lung cancer

H. C. Donker[1], B. van Es[2,3], M. Tamminga[1,4], G. A. Lunter[5], L. C. L. T. van Kempen[6], E. Schuuring[6], T. J. N. Hiltermann[1], and H. J. M. Groen[1]

[1] *Department of Pulmonary Diseases, University of Groningen, University Medical Centre Groningen, Hanzeplein 1, p.o. box 30.001, 9700 RB Groningen, the Netherlands*

[2] *Central Diagnostic Laboratory, University Medical Centre Utrecht, Utrecht University, Heidelberglaan 100, 3508 GA Utrecht, The Netherlands*

[3] *MedxAI, Theophile de Bockstraat 77-1, 1058VA Amsterdam, the Netherlands.*

[4] *Department of Internal Medicine, Twente Hospital, Netherlands*

[5] *Department of Epidemiology, University Medical Centre Groningen, 9713 GZ Groningen, The Netherlands*

[6] *Department of Pathology and Medical Biology, University Medical Centre Groningen, Hanzeplein 1, p.o. box 30.001, 9700 RB Groningen, the Netherlands*

Abstract

In advanced non-small cell lung cancer (NSCLC), response to immunotherapy is difficult to predict from pre-treatment information. Given the toxicity of immunotherapy and its financial burden on the healthcare system, we set out to identify patients for whom treatment is effective using mutational signatures from DNA mutations in pre-treatment tissue. Analysis of single base substitutions, doublet base substitutions, indels, and copy number alteration signatures in the discovery set ($m = 101$ patients) linked tobacco smoking signature (SBS4) and thiopurine chemotherapy exposure-associated signature (SBS87) to durable benefit. Combining both signatures in a machine learning model separated patients with a progression free survival hazard ratio of $0.40^{+0.28}_{-0.17}$ on the cross validated discovery set and $0.24^{+0.31}_{-0.14}$ on an independent external validation set ($m = 56$). This paper demonstrates that the fingerprints of mutagenesis, codified through mutational signatures, can be used to select advanced NSCLC patients who may benefit from immunotherapy, thus reducing unnecessary patient burden.

Introduction

In non-small cell lung cancer (NSCLC), response to immunotherapy is low, with radiology-assessed response typically around 20-25% [1], while the percentage of patients achieving durable benefit (DB), defined as progression free survival (PFS) $\geq \frac{1}{2}$ year, is only slightly higher. As a result, the majority of patients are subjected to a futile toxic treatment.

Predictors can help to narrow down specific subpopulations for which treatment is particularly effective. Programmed death ligand 1 (PD-L1) protein expression in tumor tissue, and tumor mutational burden (TMB), defined as the number of acquired amino acid sequence-changing mutations [2], is currently used to predict efficacy of immunotherapy in NSCLC [3]. The explanation behind the predictive value of TMB is that the accumulation of mutations in coding DNA results in a high diversity of neoantigens, which induces a broad anti-tumor adaptive immune response. *In silico* analysis of neoantigens found a near-perfect relation between TMB and inferred number of neoantigens [4]. Once the cancer cell expresses neoantigens, the cancer cell can be eliminated through immune recognition and cell killing [1, 2].

While several studies show a clear association of TMB with response to treatment, not all do [5], motivating the search for an improved proxy for treatment efficacy. The chief advantage of TMB is that it is easy to compute by pooling all amino acid sequence-changing mutations from whole-exome sequencing data [2], irrespective of their genomic context. However, this ignores the fact that somatic mutations are generated by a range of processes. Nucleotide context is a clue to the genesis of mutations [6], and analysis of mutation spectra, which partition mutations by alteration and DNA context, have revealed that both exogenous and endogenous DNA mutational processes can be linked to specific signatures [7, 8, 9]. Over the last decade, increasingly large datasets such as the COSMIC database have enabled the systematic identification of mutational signatures and in many cases the elucidation of their aetiology, particularly for single base substitutions [10, 11]. For instance, specific DNA mutational signatures or "genomic scars" left by e.g. polycyclic aromatic hydrocarbons [12], ultraviolet light exposure [13], and platinum chemotherapy [14] have all been validated in experiments.

Evidence is accumulating that mutational signatures in cancer are therapeutically relevant [9]. Given the partial success of TMB for predicting immunotherapy efficacy in NSCLC [2], a logical next step is to try and improve TMB by sieving out irrelevant mutations. Mutational signatures, which disentangle mutations by their proposed root cause, may help to pinpoint relevant alterations. In fact, specific mutational signatures (e.g., APOBEC A3A) have been hypothesised to be promising candidates for immune stimulation treatments [15]. Along this line, previous studies performed *de novo* identification of single base substitution signatures in NSCLC and linked these to immunotherapy efficacy [16, 17]. Given the relatively small datasets used in these studies, only three signatures were identified in both cases [16, 17]. Specifically, Wang *et al.* [16] found that patients with durable clinical benefit were enriched in signatures with similari-

ties to COSMIC signatures SBS2 and SBS13 that are associated with damage from APOBEC, a family of enzymes that are part of the innate anti-retroviral defense that operate by generating mutations in single-stranded DNA [9, 18]; and signatures similar to clock-like signature SBS1, capturing mutations that steadily accrue with age, were linked to non-response to immunotherapy by Chong *et al.* [17].

Here, instead of *de novo* analysis we directly use previously catalogued mutational signatures like in earlier work [4, 19]. We expand previous efforts by interrogating an order of magnitude more signatures, including the recently developed copy number signatures [20, 21], for their role in eliciting an immune response, and validate the presence of single base substitutions signatures at the RNA level. The primary goal of this work is to develop and validate an immunotherapy efficacy model for advanced NSCLC patients to help reduce ineffective treatment. We hypothesise that the genomic scars in pre-treatment tumor tissue, decomposed into mutational signatures, can identify durable immunotherapy beneficiaries (Fig. 1a).

Result & Discussion

In total, $M = 157$ patients with advanced NSCLC were analysed, $m = 101$ in the discovery and $m = 56$ in a separate validation set. The population had a mean age of 63.2, consisted of an approximately equal split between males and females (53.5%), and the majority of the patients (85/147, 57.8%) did not achieve durable benefit from immunotherapy (Table 1).

Mutational signatures in the discovery set.

Whole genome sequencing revealed a mean of 7.80 Mb^{-1} [median 5.43 Mb^{-1}; inter quantile range (IQR) 3.90-9.85 Mb^{-1}] non-synonymous variants in the discovery set, consistent with previous reports on whole exome sequencing [22, 23].

Mutational signature data of $p = 47$ single base substitutions (SBS), $p = 11$ doublet base substitution (DBS), $p = 17$ short insertion and deletion (indel), and $p = 20$ copy number signatures were determined (see Fig. 1b for a schematic overview) and analysed, after discarding sequencing artefact attributed signatures (Supplementary Table 1). Ranking mutational signature attribution $\boldsymbol{W}$ (see the Methods for details) by the median (across samples) shows that nine SBS signatures are present in the majority of samples (Fig. S2a, Supplementary Material), with signature SBS4, a highly lung cancer-specific signature [24] that is linked to tobacco smoking [6, 7, 12], having the highest median value (0.91 Mb^{-1}), although exhibiting considerable variability (IQR: 0.00-3.14 Mb^{-1}). Signature attributions of indel and doublet base substitution were an order of magnitude lower (Fig. S2b, Supplementary Material) with the two highest median signatures, ID3 [7, 10] (median: 0.11 Mb^{-1}, mean: 0.17 Mb^{-1}) and DBS2 [7, 25] (median: 0.06 Mb^{-1}, mean: 0.10 Mb^{-1}), both attributed to tobacco smoking,

Table 1: Patient characteristics and outcome of advanced non-small cell lung cancer. Symbols and abbreviations: αCTLA-4, cytotoxic T lymphocyte-associated antigen-4 inhibitor; αPD-1, programmed death-1 inhibitor; αPD-L1, programmed death-ligand 1 inhibitor; αTNFRSF7, tumor necrosis factor receptor superfamily type 7 inhibitor; αVEGF-A, vascular endothelial growth factor inhibitor; μ, mean; σ, standard deviation; DR, data request; Q_1, first quartile; Q_3, third quartile.

		Discovery (Hartwig DR #094)	**Validation** [Miao *et al.* ('18)]	**Overall**
m		101	56	157
Age, μ (σ)		63.8 (8.7)	61.5 (8.7)*	63.2 (8.8)
Gender, m (%)	female	52 (51.5)	32 (57.1)	84 (53.5)
Smoker, m (%)	current		14 (25.0)	14 (8.9)
	former		29 (51.8)	29 (18.5)
	never		13 (23.2)	13 (8.3)
	unknown	101 (100.0)		101 (64.3)
Prior therapy, m (%)	chemotherapy	33 (32.7)		33 (21.0)
	naive	18 (17.8)		18 (11.5)
	radiotherapy	7 (6.9)		7 (4.5)
	radiotherapy + chemotherapy	43 (42.6)		43 (27.4)
	unknown		56 (100.0)	56 (35.7)
Treatment, m (%)	αCTLA-4 + αPD-1	1 (1.0)		1 (0.6)
	αCTLA-4 + αPD-1 with chemotherapy	1 (1.0)		1 (0.6)
	αPD-1	77 (76.2)		77 (49.0)
	αPD-1 with chemotherapy	13 (12.9)		13 (8.3)
	αPD-1/αPD-L1		56 (100.0)	56 (35.7)
	αPD-L1	3 (3.0)		3 (1.9)
	αPD-L1 with chemotherapy	2 (2.0)		2 (1.3)
	αPD-L1/cabozantinib	1 (1.0)		1 (0.6)
	αTNFRSF7 + αPD-1	1 (1.0)		1 (0.6)
	αVEGF-A + αPD-1	2 (2.0)		2 (1.3)
Durable benefit, m (%)	no	57 (56.4)	28 (50.0)	85 (54.1)
	yes	36 (35.6)	26 (46.4)	62 (39.5)
	unknown	8 (7.9)	2 (3.6)	10 (6.4)
Sequencing, m (%)	whole exome (DNA)		56 (100.0)	56 (35.7)
	whole genome (DNA)	101 (100.0)		101 (64.3)
	paired whole transcriptome (RNA)	36 (35.6)		36 (22.9)
Tissue, m (%)	formalin-fixed paraffin-embedded		56 (100.0)	56 (35.7)
	fresh frozen	101 (100.0)		101 (64.3)
Tumor content sample, median [Q_1, Q_3]		0.4 [0.3,0.6]	0.3 [0.2,0.5]	0.4 [0.3,0.6]

* Age was missing for 15 patients in the validation cohort.

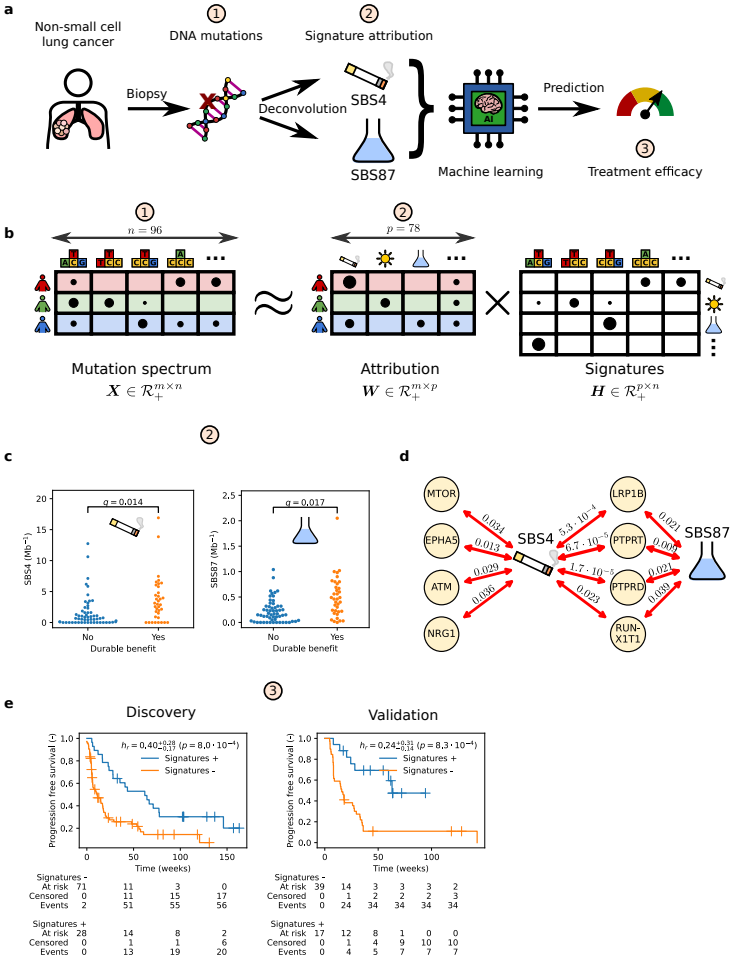


Figure 1: **Mutational signatures from pre-treatment tumor tissue predict immunotherapy efficacy in advanced non-small cell lung cancer.** **a**, Single base substitutions (SBS) are determined from pre-immunotherapy tumor material. After deconvolution into signature attributions, a machine learning classifier uses smoking-associated signature SBS4 and thiopurine chemotherapy-associated signature SBS87 to predict durable benefit (DB) from immunotherapy. **b**, Cartoon illustration of SBS signature deconvolution, where we solve for signature attribution $\boldsymbol{W}$ given mutation spectrum $\boldsymbol{X}$ and COSMIC signatures $\boldsymbol{H}$ through $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$. Nucleotide pyramids indicate SBS with flanking context; Sun, cigarette, and Erlenmeyer symbols depict example aetiologies; Shading highlights information that pertains to the corresponding patient. For illustration purposes, the size of the dots do not represent actual data. **c**, Signatures SBS4 [$q = 0.014$, Benjamini-Hochberg corrected Kolmogorov-Smirnov (B-HK-S) test] and SBS87 ($q = 0.017$, B-HK-S test) are linked to DB (discovery set). **d**, Signatures SBS4 and SBS87 correlate with mutations in genes canonically mutated in cancer (discovery set). Correlations were assessed using a B-H corrected Kendall τ (corrected p-values along the arrows) and the correlation strengths are indicated by the arrow line widths. **e**, Patients predicted to have DB (Signatures +, blue line) have superior progression free survival compared to those predicted to have non-DB (Signatures -, orange line) in the discovery set (left panel). The classifier's performance replicates in an independent validation set (right). Censored observations are indicated by crosses. Estimates and corresponding 95% confidence intervals are indicated by sub and superscripts.

like SBS4. Whole genome copy number deconvolution revealed that homologous recombination deficiency [21] signature CN17 had the highest median attribution (median: 37.2, mean: 46.5, Fig. S2c, Supplementary Material).

Mutational signatures linked to durable benefit.

Next, we looked for pre-immunotherapy mutational signatures that were determinants of therapy efficacy. Univariate analysis ($m = 93$ patients) singled out two single base substitution signatures (Fig. 1c). Tobacco smoking signature SBS4 was significantly different ($q = 0.014$, B-HK-S test) in patients who derive DB from immunotherapy. This finding underpins earlier work that found that smoking attributed transversion-high tumors [26], or an enrichment in smoking signature [19], had improved outcome in ICI treated NSCLC. Similarly, SBS87—whose mutations coincides with thiopurine chemotherapy exposure [9, 27]—was also found to differ between both groups of patients ($q = 0.017$, B-HK-S test). Note that, according to the clinical records available to us, none of the patients have been treated with thiopurine related compounds. Both signatures were uncorrelated after normalising by the total number of amino-acid changing mutations (i.e., TMB) and stratifying by target label (Kendall $\tau = -3.6 \cdot 10^{-2}$, $p = 0.76$ and $\tau = -8.2 \cdot 10^{-3}$, $p = 0.95$ for DB and non-DB, respectively).

Earlier work linked clock-like mutational signature SBS1—capturing substitutions that steadily accrue with age—with non-response and worse survival [17]; we could not replicate the association with DB, even without multiple testing correction ($p = 0.36$, K-S test). Another study linked mutational signatures associated with APOBEC—a family of enzymes that are part of the innate anti-retroviral defense that operate by generating mutations in single-stranded DNA [9, 18]—with improved immunotherapy outcome [16]. Validation of APOBEC mutational signatures SBS2 and SBS13 [16] showed a trend for SBS13 ($p = 0.04$ K-S test, $q = 0.76$ B-HK-S test) but not for SBS2 ($p = 0.66$ K-S test, $q = 1.0$ B-HK-S test). Given that APOBEC mutagenesis is highly transient, with episodic bursts of mutations [28], our samples, that represent a snapshot in time, are perhaps less suited to fully interrogate the relevance of this mutational signature on treatment outcome. None of the doublet, indel, and genome wide copy number alteration signatures were significantly associated with DB.

Relation of signature specific mutations to specific genomic loci.

Neither the SBS4 nor the SBS87 signature was correlated to the number of mutations in any of the 23 significantly lung cancer mutated genes in TCGA [22, 23] (Supplementary Table 2). Expanding the search from 23 to the top 2.5%, 5% and 10% highly expressed genes also found no correlation. In contrast, analysis of mutations in 523 genes contained in the clinically relevant TSO500 panel, consisting of genes canonically mutated in cancer, yielded eight genes [*ATM* ($q = 0.029$), *EPHA5* ($q = 0.013$), *LRP1* ($q = 5.3 \cdot 10^{-4}$), *MTOR* ($q = 0.034$), *NRG1* ($q = 0.036$), *PTPRD* ($q = 1.7 \cdot 10^{-5}$), *PTPRT* ($q = 6.7 \cdot 10^{-5}$), *RUNX1T1*

($q = 0.023$), Kendall τ correlation test] in which mutation count correlated significantly with smoking signature SBS4 (Fig. 1d) after multiple testing correction and exclusion of non-mutated genes. When combined, the mutation count was also directly linked to DB ($p = 0.0050$, K-S test) in addition to the indirect correlation through SBS4. Four of these genes [namely, *LRP1* ($q = 0.021$), *PTPRD* ($q = 0.021$), *PTPRT* ($q = 0.0091$), and *RUNX1T1* ($q = 0.039$), but no other genes] also correlated with SBS87. Both *ATM* and *EPHA5* interact at the site of DNA repair. Adding ATM to a DB logistic regression model with TMB changed the regression coefficient by more than 10 % (0.156 versus 0.117) indicating that ATM (but not EPHA5) potential confounds TMB, although its size of $\sim$ 150 kb [29] is fairly large. *MTOR* regulates cellular metabolism and the others are tumor suppressor genes. They are involved in cell interactions such as the *PTPRT* and suppression of inflammatory responses such as *STAT3*. *PTPRD* show deleterious mutations in 9% of lung cancers. *NRG1* suppresses the transcription of inflammatory cytokines and was the only gene (out of all eight) that was significant ($p = 0.020$) when added to a logistic regression model with TMB. Compared to TMB, mutations in NRG1 anti-correlated with DB (as indicated by the negative regression coefficient -2.6) suggesting that mutations in this gene may disrupt the adaptive immune system's capacity to induce an immune response. SBS87 only affects tumor suppressor genes explaining why driver mutations are not the driving force of the immune response. However, we found no enrichment for tumor suppressor or oncogenes in the set correlating with SBS4 ($p = 0.25$ and $p = 0.16$, respectively, Fisher exact test) nor with SBS87 ($p = 0.08$ and $p = 0.16$, respectively, Fisher exact test) relative to the TSO500 gene set.

Expression of SBS4/SBS87 mutations is not a sufficient condition for predicting durable benefit.

Next, we aimed to explain the predictiveness of SBS4 and SBS87 attributed mutations by looking for differences at the RNA level. Assume that immune recognition, through the presentation of mutated protein fragments on the major histocompatability complex, is a prerequisite for eliciting an immune response. If transcription is a necessary condition for immune recognition of mutated DNA then, by extension, it must also be a necessary condition for predicting DB, assuming immunotherapy operates through an adaptive immune response. Focussing on transcripts with a SBS4 and/or SBS87 dominant variant (a signature-dominant variant accounts for $\geq 50\%$ of the signature's mutations, see Sec. A.4, Supplementary Material), we asked whether their transcription is also a sufficient condition for predicting DB.

In a subanalysis of paired RNA and DNA samples ($m = 36$ patients), durable benefit ($p = 0.67$, Fisher exact test) and both SBS4 and SBS87 signatures ($p = 0.30$, $p = 0.28$ respectively, K-S test) did not differ from the entire discovery population. Comparing the amount of transcripts containing a SBS4 or SBS87-dominant singlet versus transcripts containing any other singlet shows that the RNA abundance is similar in both groups ($p = 0.71$, K-S test, Fig. S3a,

Supplementary Material). Compared to DB, the distribution of transcripts harboring a smoking associated signature SBS4-dominant variant was different from patients with non-DB ($p = 0.0036$, K-S, see Fig. S3b, Supplementary Material). However, this difference could be attributed to the number of SBS4 DNA mutations (Fig. S3c, Supplementary Material). For the thiopurine chemotherapy associated SBS87 signature, no difference in absolute number ($p = 0.11$, K-S, Fig. S3d, Supplementary Material) nor relative to the number of corresponding variants ($p = 0.48$, K-S, Fig. S3e, Supplementary Material) was found in the amount of mutated transcripts between patients with and without DB. Differential gene expression of the mutated RNA of the aforementioned eight significant genes ($m = 22$ patients with paired DNA, RNA, and ≥ 1 mutations in any of these eight genes) detected no difference in the amount mutated RNA between patients with and without DB ($q > 0.05$ for all genes, B-HK-S test). Together, these results show no evidence that distinguishes SBS4 and SBS87 from other mutations at the transcription level in durable beneficiaries. While these negative findings could point to the differences (i) being manifested further downstream immune processes or (ii) the capacity to bind to a broader spectrum of HLA alleles [30], we do not rule out that we lacked insufficient variant coverage at the RNA level (which was on average 19.8× for all and 19.1× for SBS4/SBS87 dominant variants) to detect subtle differences.

A signature-based classifier predicts immunotherapy benefit.

In combining SBS4 and SBS87 signature attributions (representing, per signature, the number of amino-acid sequence changing singlets in DNA) both remained significant in each cross validated fold, confirming that the aforementioned univariate analysis does not lead to overfitting on the discovery set. The classifier scored an area under the receiver operating characteristic curve (ROC AUC) of $0.74^{+0.11}_{-0.12}$ on the hold-out folds (Fig. S4, Supplementary Material). This was significantly higher than when the classifier was trained on TMB (ROC AUC: $0.65^{+0.12}_{-0.13}$, $p = 0.016$ PPT, Fig. S4a, Supplementary Material). The ROC curves of individual signatures were similar to that of the model (Fig. S5). With an estimated 43.8% patients with DB [31] as our classification probability threshold, a sensitivity of $0.56^{+0.16}_{-0.16}$, a specificity of $0.86^{+0.08}_{-0.10}$, and an accuracy of $0.74^{+0.09}_{-0.10}$ was achieved (Table 2) with 8 false positive and 16 false negative classifications (Supplementary Table 3). The classifier was not perfectly calibrated (Fig. S6a), as expected from the conditional independence assumption of the naive Bayes model.

To incorporate patients censored prior to the $\frac{1}{2}$ year mark, we compared predicted versus actual outcome using Kaplan-Meier (Fig. 1e) in all $m = 101$ discovery patients. Durable benefit predicted patients had superior median progression free survival (62 for predicted DB versus 11 weeks for predicted non-DB). Cox regression provided additional confirmation that the classifier significantly predicted outcome based on mutational signatures SBS4 and SBS87

Table 2: Durable benefit (DB) classifier performance metrics and p-value comparing the performance between the two datasets. Abbreviations: ROC AUC, area under the receiver operating characteristic curve; AP, average precision; h_r, hazard ratio of progression free survival (PFS) comparing patients predicted DB versus predicted non-DB. Estimates and corresponding 95% confidence intervals are indicated by sub and superscripts.

	ROC AUC	**AP**	F_1	**Sensitivity**	**Specificity**	**Accuracy**	h_r*
Discovery							
(Hartwig DR #094)	$0.74^{+0.11}_{-0.10}$	$0.63^{+0.18}_{-0.15}$	$0.63^{+0.14}_{-0.16}$	$0.56^{+0.16}_{-0.17}$	$0.86^{+0.09}_{-0.10}$	$0.74^{+0.09}_{-0.10}$	$0.40^{+0.28}_{-0.17}$
Validation							
(Miao *et al.* ('18))	$0.69^{+0.13}_{-0.15}$	$0.71^{+0.16}_{-0.19}$	$0.57^{+0.17}_{-0.19}$	$0.46^{+0.21}_{-0.21}$	$0.86^{+0.11}_{-0.14}$	$0.67^{+0.13}_{-0.13}$	$0.24^{+0.31}_{-0.14}$
p	0.57	0.56	0.68	0.54	1.00	0.35	0.32

$*$ All patients were included in the PFS analysis. For all other classification metrics, only patients where censoring permitted unambiguous DB label assignment were analysed.

combined (hazard ratio $h_r = 0.40^{+0.28}_{-0.17}$, $p = 8.0 \cdot 10^{-4}$).

Independent validation on the data from Miao *et al.* [4] reproduces the classifier's performance. The median survival in the external validation set ($m = 56$ patients) was 63 versus 16 weeks for DB predicted patients compared to the others (Fig. 1e). All performance metrics were similar (Table 2), highlighting the reproducibility of our approach despite differences in (i) tissue handling (formalin-fixed paraffin-embedded versus fresh frozen), (ii) chemistry (whole exome versus genome capture), and (iii) bioinformatics pipeline. Unlike the discovery set, the ROC AUC of a model trained on TMB was not significantly different from the mutational signature model (ROC AUC $0.78^{+0.12}_{-0.14}$ versus $0.69^{+0.14}_{-0.14}$, respectively, $p = 0.18$ PPT, Fig. S4b, Supplementary Material). This discrepancy was attributed to a difference in TMB distribution between the discovery and validation set ($p = 0.031$, KS test), while both SBS4 and SBS87 signature attributions remained similar in both cohorts ($p = 0.082$ and $p = 0.69$, respectively, KS test). A fair head-to-head comparison between TMB and the mutational signature approach requires a separate, substantially larger, study with more detailed patient characteristics and medical history than presented here (together with a more harmonised tissue handling, sequencing chemistry, and upstream bioinformatics preprocessing).

Error analysis of the discovery set revealed that treatment naive patients were over represented in the top ten worst predicted false negatives ($p = 0.015$, Fisher exact test). Exclusion of ($m = 18$) treatment-naive patients slightly improved the model (ROC AUC: $0.79^{+0.11}_{-0.12}$) on the discovery set. More strikingly, after training on pre-treated patients only, generalization on the validation patients—of whom prior therapy was unknown—also improved (ROC AUC: $0.71^{+0.14}_{-0.15}$). Caution is therefore warranted, when applying the classifier to a (in our study, underrepresented) treatment naive population.

Subanalysis of patients with smoking status ($m = 54$, validation set only, Table 1) showed that never smokers were more difficult to classify (ROC AUC:

$0.40^{+0.10}_{-0.15}$ versus $0.71^{+0.12}_{-0.13}$, p = 0.048 UPT), although the numbers were low with only $m = 3$ durable beneficiaries in the never-smokers group. Note that current/former smokers are known to have a better overall response rate [32]. Seeing that the performance was lower on the combined set indicates that the mutational signatures provide information that is orthogonal (or, complementary) to smoking status.

Decision curve analysis.

Net benefit is a decision-theoretic concept that quantifies the practical utility of a classifier [33, 34]. The integrated net benefit of using the signature based model is positive (Figs. S7 and S8, Supplementary Material), with a median integrated combined net benefit (see e.g. Talluri *et al.* [35]) of 0.37 (IQR: 0.366-0.379) and 0.29 (IQR: 0.284-0.296) with respect to net benefits less than zero when treating *all* patients for the discovery and validation datasets, respectively. In general, according to Ref. [34] a model can be recommended for clinical use if, across a range of clinically reasonable probability thresholds, it has the highest level of benefit. This range of thresholds can be viewed as the acceptable range of *number-needed-to-treat* (NNT) to have one effective treatment and is clinician/protocol dependent. In Figs. S7 and S8, we see a net-benefit for both the treated/untreated with respect to treating all patients for a probability threshold-range of roughly $0.3 - 0.6$ and $0.4 - 0.6$ for the discovery and validation datasets respectively. Important to note, although we have fairly poor calibration for the $0.3 - 0.6$ range within which we expect a relative net benefit with respect to the baseline, outside this range the net benefit is roughly equal. That is, within a probability threshold range of $0.3 - 0.6$, so an NNT-range of roughly $1.5 - 3$ patients, we have a net benefit when applying our model, outside this range there is no added benefit.

Conclusion

We have identified tobacco smoking signature SBS4 and the recently identified thiopurine chemotherapy exposure-associated signature SBS87 [9, 27] as factors that are predictive of benefit from immunotherapy. Both signatures are linked to mutations in genes involved in DNA repair and/or the function of tumor suppressor genes that have a role in immunological cellular interactions and inflammatory cytokines. In contrast, none of the doublet base substitution, indel, or copy number alteration signatures were associated with durable benefit from immunotherapy. RNA analysis of the two signatures found no evidence that distinguishes these from other mutations in terms of immunotherapy efficacy.

When combined, these two signatures can help to select advanced NSCLC patients who benefit from immunotherapy using information available prior to treatment initiation. The advantage of our approach is that it inherits the mechanistic grounding of mutational signatures and helps to elucidate the underlying immunological mechanisms that predict durable benefit. However, more

research is needed to establish if our approach works sufficiently well on the (underrepresented) treatment-naive and non-smoking patient populations, or if additional adjustments are needed.

Online Methods

Cohort assembly

We retrospectively compiled clinical records and data of tumor and matched normal tissue (collected prior to treatment initiation) of immunotherapy treated advanced NSCLC patients. To this end, patient characteristics, whole genome sequencing (WGS) and total RNA data derived from fresh frozen biopsies were requested from Hartwig Medical Foundation (HMF) [36]. This cohort, containing metastatic NSCLC patients, formed the discovery dataset. The external validation was extracted from Ref. [4] and consisted of whole exome sequencing (WES) of formalin-fixed paraffin-embedded (pre-immunotherapy) NSCLC samples of tumors and matched normal tissue. A more detailed description of cohort assembly can be found in the Supplementary Material.

DNA processing

Whole genome sequencing, variant calling, and purity estimation were performed by HMF [36, 37]. Since the whole genome sequencing reads were mapped to GRCh37, we used crossmap [38] to perform a liftover (or, remapping) from GRCh37 to GRCh38 using Ensembl's corresponding chain file.

Mutation deconvolution

COSMIC v3.3 (June 2022) mutational signatures, $\boldsymbol{H}$, were used to deconvolute mutations (see Supplementary Table 1 for a list of analysed signatures). For substitutions and short indels, these signatures $\boldsymbol{H}$ describe the nucleotide alteration distribution [10, 39]. Release v3.3 adds the recently developed copy number signatures which captures the copy number $\times$ zygosity $\times$ length distribution [20, 21].

We first used SigProfilerMatrixGenerator on amino acid-changing mutations to extract mutation spectra of single base substitutions (singlets) with two flanking bases (SBS-96), doublet base substitutions (doublets, DB-78), and insertion deletions (indels, ID-83) [40]. The same package was used to compute genome-wide copy number alterations (CN-48) [21]. Each mutation spectrum, $\boldsymbol{X}$, is a positive m-by-n matrix (i.e., $\boldsymbol{X} \in \mathcal{R}_{+}^{m \times n}$) consisting of m samples, and n mutation/channels ($n = 96$, $n = 78$, $n = 83$, $n = 48$ for SBS-96, DB-78, ID-83, and CN-48, respectively) counting the number of mutations per channel. The positive mutational signature matrix $\boldsymbol{H}$ relates the p signatures (rows) to the corresponding n mutation type/channels (columns), $\boldsymbol{H} \in \mathcal{R}_{+}^{p \times n}$. (Throughout this paper we adhere to the convention that m, n, and p indicate the number

of patients, number of mutation types/channels, and number of signatures [except for p-values which will be clear from the context], respectively.) Using the spectrum $\boldsymbol{X}$ and mutational signatures $\boldsymbol{H}$, signature attribution $\boldsymbol{W}$, a positive m-by-p matrix (i.e., $\boldsymbol{W} \in \mathcal{R}_{+}^{m \times p}$) such that $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$, was computed by non-negative matrix factorisation using a coordinate descent solver with an error tolerance of 10^{-6} for no more than 10^4 iterations.

Signature attributions that refer to possible sequencing artefacts [39] or with zero variance in either durable benefit stratum in the discovery set were excluded from the analysis. Mutational signatures of singlets, doublets, indels, and TMB (i.e., total mutation count) were obtained from non-synonymous mutations and normalised by (exome) coverage size in mega bases (Mb) and rounded to two decimals to retain three significant digits. For the Hartwig WGS, the exome coverage size in individual samples was unknown, and we therefore took a size of 47.9 Mb [41]. For the discovery cohort (Miao *et al.* [4]), we used the size as indicated in their Supplementary Table 1. Since the copy number variants span large portions of the genome (both exonic and intronic regions), we report the total, whole genome, copy number attributions.

RNA processing

Singlets were traced back to transcripts to study how the mutational signatures manifest at the transcription level, as a surrogate for protein expression. Out of the $m = 101$ patients in the discovery cohort, raw (total) RNA sequencing data of $m = 40$ patients were available (no RNA was available in the validation cohort). Briefly, raw sequencing data were trimmed, aligned, and converted in to transcripts per million (TPM). After quality control, two inferior quality samples and two samples with insufficient follow up were excluded, leaving a total of $m = 36$ samples for analysis. The amount of transcripts containing a variant were re-estimated to account for differences in tumor content. RNA per signature was obtained by pooling transcripts containing the $\geq 50\%$ dominant mutations of the given signature. A more detailed description of the method can be found in the Supplementary Information.

Statistical Analysis

Here and in the following, all tests were two-sided.

Univariate analysis

Difference in mutational signature distributions were determined by a Kolmogorov-Smirnov (K-S) test. When more than one mutational signature was considered at a time, the Benjamini-Hochberg (B-H) correction was applied to control for false positive discoveries. Correlations between signatures and mutation counts per gene were evaluated using Kendall τ rank correlation (with B-H correction) to account for ties (both measures were derived from amino-acid sequence

changing mutations only). Significantly correlated genes were subsequently annotated as (i) tumor suppressor gene using the tumor suppressor gene database website version 2.0 [42] (accessed 1st September, 2022) and (ii) as oncogenes when present in the Cancer Gene Census COSMIC v.96 [43] (accessed 16th September, 2022) and tested for enrichment. For differential gene expression on the transcripts with a variant, we used a non-parametric B-HK-S test because for some genes no mutated transcripts were measured (these zeros were not possible to analyse with DESeq2). We use q to denote the multiple testing corrected p-values and a significance level $\leq 5\%$ was considered statistically significant.

Efficacy classifier

Patients were labelled as durable benefit [progression free survival (PFS) $\geq \frac{1}{2}$ year] or non-durable benefit (PFS $< \frac{1}{2}$ year), whenever censoring permitted unambiguous label assignment. All patients were included for PFS analysis. To predict outcome, we used a naive Bayes classifier: a classic supervised machine learning method [44, 45] that works particularly well with few samples, even when its conditional independence assumption is violated [46]. Features were modelled with a zero-inflated exponential distribution. Results reported on the discovery cohort were obtained by leave-one-out cross validation while inference on the validation set was done after training on the entire discovery cohort.

Estimates a and ninety five per cent confidence intervals $[a - b, a + c]$ of the average precision, area under the receiver operating characteristic curve (ROC AUC), F_1 score, sensitivity, and specificity were estimated by bootstrapping for 1000 iterations and denoted as a^{+c}_{-b}. Head-to-head model comparison were evaluated using a paired permutation test (PPT), while performance estimates of different sets were compared using an unpaired permutation test (UPT).

For the PFS analysis (including all patients), we visualised the (out-of-fold) predicted versus actual PFS outcome using the Kaplan-Meier method. To quantify agreement, hazard ratios and significance were evaluated using Cox regression (for which we tested appropriateness). To compare hazard ratios h_r, corresponding coefficients (i.e., $-\ln h_r$), were compared with a regression coefficient test [47].

Funding

This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904.

Acknowledgements

We thank Fenneke Zwierenga en Benthe Muntinghe for proofreading. This publication and the underlying research are partly facilitated by Hartwig Medical Foundation and the Center for Personalized Cancer Treatment (CPCT) which have generated, analysed and made available data for this research.

Data availability statement

The discovery set can be requested from Hartwig Medical Foundation, data request # 094. The validation set can be extracted from Ref. [4]. Code and analysis notebooks are publicly available on Gitlab: `https://gitlab.com/hylkedonker/genomic-scars-predictor-nsclc-immunotherapy` under the MIT license.

Conflicts of interest

HCD: None to declare; BvE: None to declare; MT: None to declare; GAL: None to declare; ES: Honoraria/speakers fee: Bio-Rad, Roche, Agena Bioscience, Illumina, Lilly; Consulting or Advisory Role: MSD/Merck, Astellas, Bayer, BMS, Agena Bioscience, Janssen Cilag (Johnson & Johnson), Novartis, Roche, AstraZeneca, Amgen, Lilly; Research Funding: Biocartis, Bio-Rad, Roche, Agena Bioscience, AstraZeneca, InVitae/Archer (all paid to UMCG); Travel, Accommodations, Expenses: Roche Molecular Diagnostics, Bio-Rad. LCLTvK: Grants, non-financial support from Roche, advisory board presence for AstraZeneca, Novartis, Merck, Janssen-Cilag, Bayer, BMS, nanoString and Pfizer, grants and non-financial support from Invitae, non-financial support from Biocartis, grants from Bayer, non-financial support from nanoString. TJNH: Advisory/consultancy fees from AstraZeneca, Bristol-Myers-Squibb, Merck Sharp Dohme, Roche, and research grants/funding from AstraZeneca, Hoffmann-La Roche. HJMG: Consulting or Advisory Role: Novartis, Lilly, Roche/Genentech.

Author contribution

Conceptualization: HCD, BvE, MT, GAL, HJMG; Methodology: HCD; Software: HCD; Validation: HCD; Formal analysis: HCD, BvE; Investigation: HCD; Data Curation: HCD, BvE; Writing – Original Draft: HCD; Writing – Review & Editing: HCD, BvE, MT, GAL, LCLTvK, ES, TJNH, HJMG; Visualization: HCD, BvE, TJNH; Supervision: HJMG; Funding acquisition: HCD, BvE, MT, HJMG.

Supporting Material

A Extended Methods

A.1 Cohort Assembly

Patient characteristics and biopsy data were requested from Hartwig Medical Foundation (HMF), data request number 094. The HMF dataset was used as the discovery cohort and contains samples from patients with advanced cancer not curable by local treatment that were recruited from 41 hospitals in the Netherlands [36]. Limited survival data and no performance score, ethnic background, nor smoking status were available for analysis. Whenever available, response data (of unspecified evaluation criterion) were used to construct progression free survival data. Eight patients without response measurement who died shortly after treatment stop were marked as progressive disease at the last follow-up. After excluding seven patients without DNA sample and one patient that was not treated with immunotherapy, we compiled a dataset with a total of $m = 101$ patients. Accrual of these patients was between January 2016 and May 2019, with the last follow-up information from May 2020. Whole DNA was extracted from freshly frozen biopsies (of unspecified histology), and total RNA isolation was performed as described previously [36].

The external validation was extracted from Ref. [4] and consisted of whole exome sequencing (WES) of formalin-fixed paraffin-embedded pre-treatment samples from immune checkpoint inhibition treated advanced lung cancer. In turn, variants and clinical characteristics in that work consisted of newly sequenced samples combined with data from Refs. [26, 48] (all advanced lung cancer). One patient, who had small cell instead of non-small cell lung cancer, was excluded leaving $m = 56$ patients in total. No study dates were specified in Ref. [4]. No RNA was available for analysis in the validation set.

A.2 Upstream RNA processing

FASTQ files were processed by sample-wise pooling of reads from different lanes and subsequent trimming with Trimmomatic [49] using a four-base sliding window to cut reads falling below a quality of 25, dropping reads shorter than 50 bases, and trimming the 8 leading bases (the head) if below the quality threshold (and otherwise default parameters, as suggested in the manual). The subsequent sequencing data were analysed using FastQC and aligned with STAR [50] using the thirteenth release for the GRCh38 human reference assembly and Gencode's gene annotation, version 36, to guide read mapping. After alignment, MultiQC [51] was used to compare overall sample quality allowing us to exclude two anomalous samples of inferior quality. Two additional RNA samples were excluded from downstream analysis because the patients had insufficient follow-up to unambiguously assign a durable benefit label (leaving a total of $m = 36$). The remaining aligned samples (in coordinate sorted BAM format) were in-

dexed using samtools [52]. The somatic DNA mutations were tracked at the RNA level by generating pile-ups at the (remapped) genomic position to obtain the variant allele frequency, ω_{mut} (un-remappable variants were discarded). The number of RNA counts were estimated per exon (as suggested in the manual) using HTSeq-count [53] with the aforementioned gene annotation. To estimate the amount of mutant RNA we first computed the number of transcripts per million, t_α, of transcript α. Briefly, using the RNA counts per transcript, $n^{(\alpha)}$, and the transcript length, $l^{(\alpha)}$, (in base pairs), its' value is defined by

$$t_\alpha = \frac{n^{(\alpha)}}{l^{(\alpha)}} \frac{10^6}{Z}, \tag{1}$$

with Z a normalisation constant $Z = \sum_\alpha \frac{n^{(\alpha)}}{l^{(\alpha)}}$.

A.3 Mutated RNA re-restimation

Observe that the amount of mutated RNA molecules depend on the sample's overall tumor purity f_t, as well as the tendency of tumor cells to express and/or break down the mutated transcripts. We therefore introduce a (transcript specific) enrichment factor, r_α, defined as the ratio between the allele frequency $\omega_{\text{mut}}^{(\alpha)}$ in transcript α and overall tumor purity, $r_\alpha = \frac{\omega_{\text{mut}}^{(\alpha)}}{f_t}$. This enrichment factor, r_α, re-estimates the RNA from the samples as if they had 100% tumor purity. When a transcript contained more than one mutation, the allele frequency was, for simplicity, averaged. Finally, the estimated the amount of mutant RNA molecules is computed as:

$$n_{\text{mut}}^{(\alpha)} = r_\alpha t_\alpha. \tag{2}$$

A.4 RNA per signature

For each mutational signature, i, dominant mutations, j, were determined by selecting the smallest set of mutations, T, that account for $\geq 50\%$ of somatic mutations (i.e., $\sum_{j \in T} H_{ij} \geq 0.5$). Using these dominant mutations, RNA expression of transcripts containing mutations in set T [Eq. (2)] were pooled for analysis.

A.5 Net benefit

For the net-benefit analysis we use the following definitions from [54] and [34]

$$\text{NB}_{\text{treated}}(t) = \frac{\text{TP}}{m} - \frac{\text{FP}}{m} \frac{t}{1-t}\,, \tag{3}$$

$$\text{NB}_{\text{untreated}}(t) = \frac{\text{TN}}{m} - \frac{\text{FN}}{m} \frac{1-t}{t}\,, \tag{4}$$

where t is the probability threshold for accepting a positive prediction, and TP, FP, TN, FN are the number of true positives, false positives, true negatives, and

false negatives, respectively. Here a low t indicates that we attach great value to avoiding false negatives, and a high t if we attach great value to avoiding false positives. The combined net benefit is defined as it's sum $\mathrm{NB}_{\mathrm{combined}}(t) = \mathrm{NB}_{\mathrm{treated}}(t) + \mathrm{NB}_{\mathrm{untreated}}(t)$. A metric that directly follows is the integrated net benefit from [35]

$$\mathrm{NBI} = \int_0^1 \mathrm{NB}(t)\mathrm{d}t\,. \tag{5}$$

A downside of the net benefit is lack of interpretability and it does not transparently include quality adjusted life years or financial cost. To augment the net benefit we also consider cost as a function $C(t)$ of true positives, true negatives, false positive and false negatives, assuming independent costs:

$$C(t) = \frac{\omega_{\mathrm{TN}}\mathrm{TN}(t) + \omega_{\mathrm{TP}}\mathrm{TP}(t) + \omega_{\mathrm{FN}}\mathrm{FN}(t) + \omega_{\mathrm{FP}}\mathrm{FP}(t)}{m\sum_i \omega_i}\,, \tag{6}$$

where $\boldsymbol{\omega}$ represents the weights. Here weights are taken as $\omega_{\mathrm{TN}} = 0$, $\omega_{\mathrm{TP}} = 100$, $\omega_{\mathrm{FN}} = 200$, $\omega_{\mathrm{FP}} = 100$; this can be translated as; we attach a cost of 0 to true negatives as we do not incur extra costs, a cost of 100 to an accurate estimation that immunotherapy is necessary, i.e. we attach a cost of 100 to an effective immunotherapy, then a cost of 300 to a false rejection of immunotherapy to account for incurring unnecessary suffering, and again a cost of 200 for an ineffective immunotherapy, the reasoning for the latter is that we have the cost of the immunotherapy plus the cost of treating the side-effects. The attribution of weights should be done with extreme care, and we want to emphasize that we merely present this as an addition to the classical net benefit analysis.

B Extended Results

B.1 Correlation between signatures and prior treatment

The majority of patients in the discovery cohort had a history of chemotherapy and/or radiotherapy. Both radiation exposure [7, 27, 55] and chemotherapy [7] are known to cause distinguishing mutations. We therefore analysed all signatures (SBS, DB, indel, and CNV) to see if that led to any significant signature differences between the exposed versus unexposed groups. For both chemotherapy and radiotherapy, no differences between the groups was observed. Our negative findings are explained by the fact that (i) lung cancer is, after melanoma, one of the cancers with the highest mutation burden [10], (ii) signatures partially overlap, (iii) our population is small, and (iv) our false discovery control is strict. Treatment-related mutations could be buried by other mutation sources generating similar mutations — c.f., the overlap of platinum chemotherapy signature SBS35 and smoking associated signature SBS4. The resulting mutational signature differences may therefore be too small to detect in our population, given our multiple testing correction.

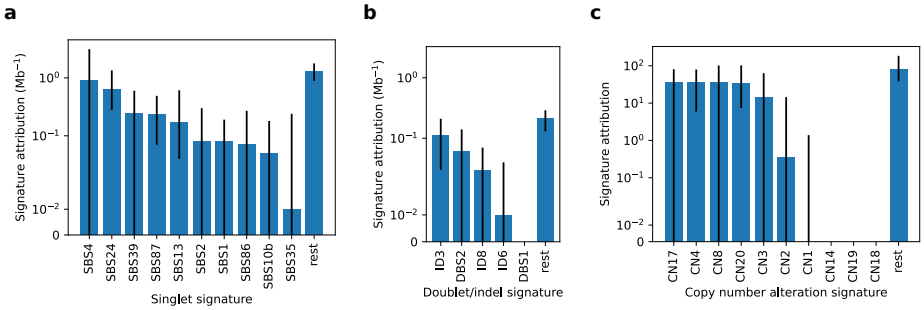


Figure S2: **Mutational signature deconvolution of tumor DNA derived from the metastatic non-small cell lung cancer discovery cohort. a**, Top ten non-synonymous single base substitution (singlet) signature attributions (prefix: SBS), per megabase (Mb) of exome. Rest: signature attributions of all other singlet signatures combined. **b**, Top five non-synonymous doublet base substitution (doublet) and short insert deletion (indel) signature attributions (prefixes DB and ID, respectively) per Mb of exome. Rest: signature attributions of all other doublet and indel signatures combined. **c**, Top ten whole genome copy number (CN) alteration signatures. Signatures are ranked by their total, genome wide, median attribution. Values and error bars indicate median and upper-lower quartiles, respectively.

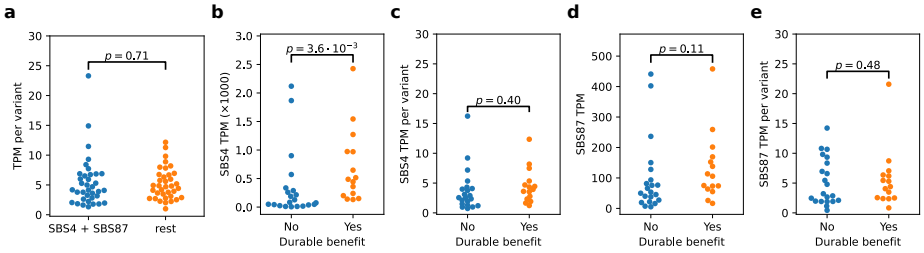
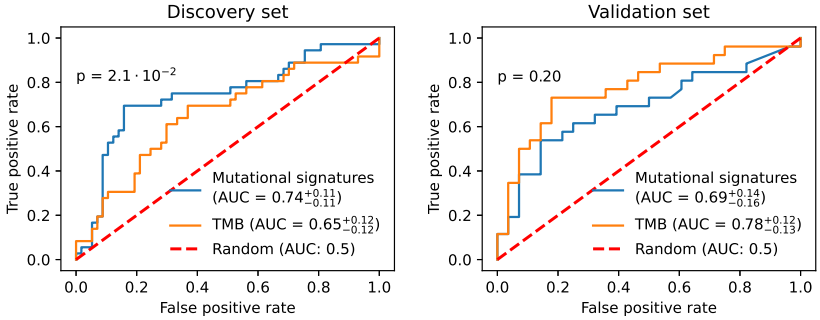


Figure S3: **Amount of mutation carrying RNA, per signature, in pre-immunotherapy tumor tissue from the discovery set. a**, RNA harboring (smoking associated) signature SBS4 or (thiopurine exposure associated) signature SBS87-dominant single base substitutions (singlets) are compared to RNA from all other singlets (rest). The amount is normalised by the number of corresponding mutations. **b**, Amount of RNA harboring signature SBS4-dominant singlets split by outcome. **c**, As in **a**, but normalised by the number corresponding mutations. **d**, Amount of RNA harboring signature SBS87-dominant singlets split by outcome. **e**, As in **c**, but normalised by the number corresponding mutations. Significance was assessed using a Kolmogorov-Smirnov test. Abbreviations: TPM, transcripts per million.



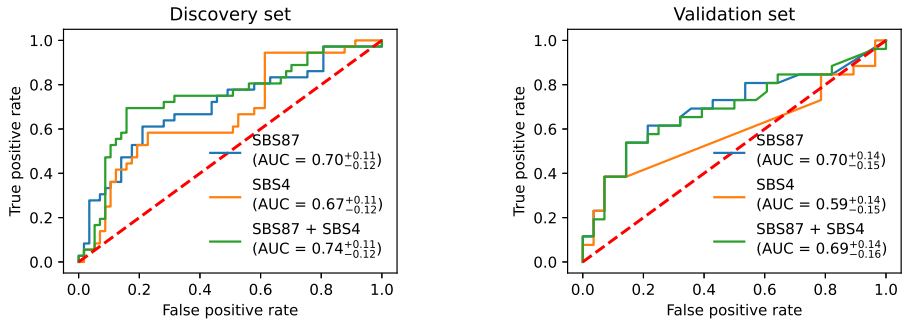
(a) Discovery dataset (Hartwig)

(b) Validation dataset [Miao *et al.* ('18)]

Figure S4: Receiver operating characteristic (ROC) curve analysis of durable benefit models that use pre-immunotherapy tumor tissue in non-small cell lung cancer. A naive Bayes classifier is trained on the discovery set of using either tumor mutational burden or with the mutational signature SBS4 and SBS87-based model. In the discovery dataset (left), the ROC area under the curve (AUC) of the mutational signature-based model is significantly higher than the TMB-based model [$0.74^{+0.10}_{-0.12}$ versus $0.65^{+0.12}_{-0.13}$, respectively, $p = 0.016$ paired permutation test (PPT)]. In the validation dataset (right), there was no significant difference in the ROC AUC ($0.69^{+0.14}_{-0.14}$ versus $0.78^{+0.12}_{-0.14}$, respectively, $p = 0.18$ PPT). Estimates and corresponding 95% CIs are indicated by sub and superscripts.

B.2 Whole genome sequencing signatures linked to durable benefit

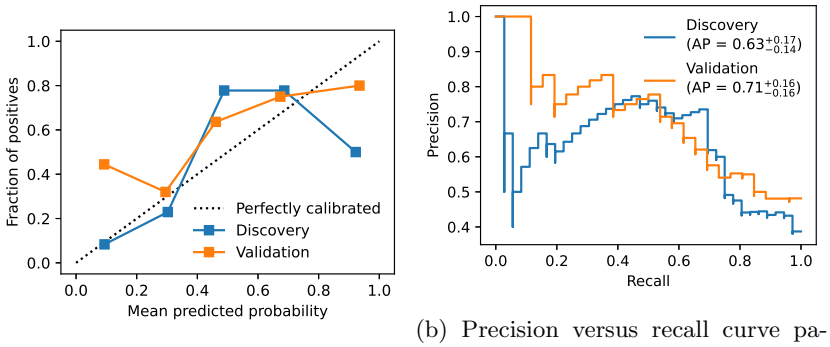
Signatures derived from all mutations (i.e., whole genome, including synonymous variants) found primarily smoking related signatures, SBS4 ($q = 0.0081$, B-HK-S test), doublet base substitution DBS2 ($q = 0.0075$, B-HK-S test), and indel signature ID3 ($q = 0.044$, B-HK-S test) associated with durable benefit. While signature SBS87 was not recapitulated in WGS ($q = 1$, $p = 1.0$), we did find another doublet signature, DBS6, of unknown aetiology associated with DB ($q = 0.0090$, B-HK-S).



(a) Discovery dataset

(b) Validation dataset

Figure S5: Receiver operating characteristics curve analysis of naive Bayes classifier with either one mutational signature, or both. Combining both signatures slightly improves the model on the discovery set (left). Holdout predictions on the external validation (right) set shows that the SBS87 classifier performs about the same as the one combining both mutational signatures. The area under the curve and 95% confidence interval are indicated in parentheses in the legend.



(a) Calibration curve as a function of classification probability.

(b) Precision versus recall curve parameterised by the classification probability threshold.

Figure S6: Performance of mutational signature model that predicts durable benefit from pre-immunotherapy tumor tissue in non-small cell lung cancer. Estimates and corresponding 95% confidence intervals are indicated by sub and superscripts. Abbreviations: AP, average precision.

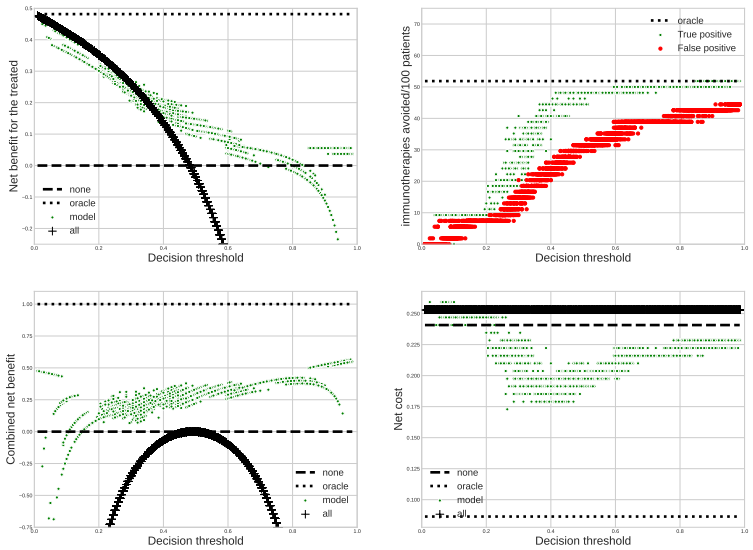


Figure S7: Net benefit plot for the discovery dataset, with three reference results, namely **none** when no patient is treated, **all** when all patients are treated and **oracle** when we have a perfect model, for the net cost curve we use $\omega_{\mathrm{TN}} = 0$, $\omega_{\mathrm{TP}} = 100$, $\omega_{\mathrm{FN}} = 300$, $\omega_{\mathrm{FP}} = 200$. The dots are obtained by resampling 50 times.

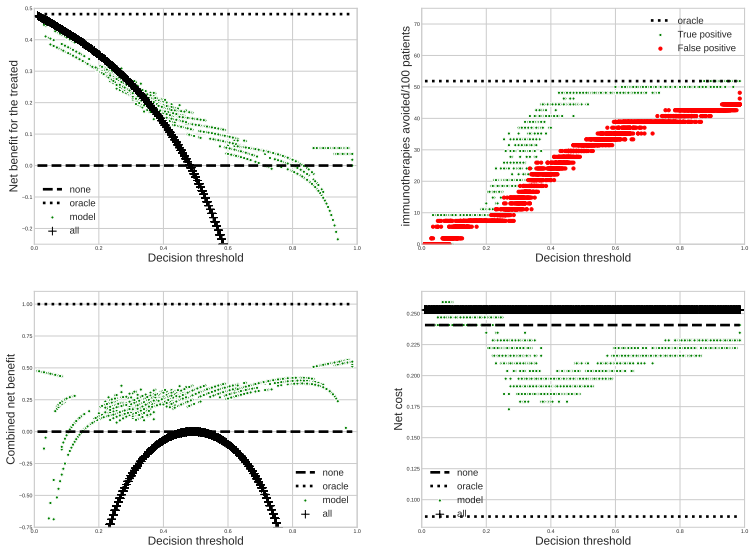


Figure S8: Net benefit plot for the validation dataset, with three reference results, namely **none** when no patient is treated, **all** when all patients are treated and **oracle** when we have a perfect model. The dots are obtained by resampling 50 times.

References

[1] D Ross Camidge, Robert C Doebele, and Keith M Kerr. "Comparing and contrasting predictive biomarkers for immunotherapy and targeted therapy of NSCLC". In: *Nature Reviews Clinical Oncology* 16.6 (2019), pp. 341–355.

[2] Lynette M Sholl et al. "The promises and challenges of tumor mutation burden as an immunotherapy biomarker: a perspective from the International Association for the Study of Lung Cancer Pathology Committee". In: *Journal of Thoracic Oncology* 15.9 (2020), pp. 1409–1424.

[3] Boris Duchemann et al. "Current and future biomarkers for outcomes with immunotherapy in non-small cell lung cancer". In: *Translational Lung Cancer Research* 10.6 (2021), p. 2937.

[4] Diana Miao et al. "Genomic correlates of response to immune checkpoint blockade in microsatellite-stable solid tumors". In: *Nature genetics* 50.9 (2018), pp. 1271–1281.

[5] LBA80Pembrolizumab Paz-Ares et al. "Pembrolizumab (pembro) plus platinum-based chemotherapy (chemo) for metastatic NSCLC: tissue TMB (tTMB) and outcomes in KEYNOTE-021, 189, and 407". In: *Annals of Oncology* 30 (2019), pp. v917–v918.

[6] Ludmil B Alexandrov et al. "Signatures of mutational processes in human cancer". In: *Nature* 500.7463 (2013), pp. 415–421.

[7] Jill E Kucab et al. "A compendium of mutational signatures of environmental agents". In: *Cell* 177.4 (2019), pp. 821–836.

[8] Yoo-Ah Kim et al. "Mutational signatures: From methods to mechanisms". In: *Annual Review of Biomedical Data Science* 4.1 (2021), pp. 189–206.

[9] Samuel W Brady, Alexander M Gout, and Jinghui Zhang. "Therapeutic and prognostic insights from the analysis of cancer mutational signatures". In: *Trends in Genetics* (2021).

[10] Ludmil B Alexandrov et al. "The repertoire of mutational signatures in human cancer". In: *Nature* 578.7793 (2020), pp. 94–101.

[11] Harald Vöhringer et al. "Learning mutational signatures and their multidimensional genomic properties with TensorSignatures". In: *Nature communications* 12.1 (2021), pp. 1–16.

[12] Serena Nik-Zainal et al. "The genome as a record of environmental exposure". In: *Mutagenesis* 30.6 (2015), pp. 763–770.

[13] Douglas E Brash. "UV signature mutations". In: *Photochemistry and photobiology* 91.1 (2015), pp. 15–26.

[14] Arnoud Boot et al. "In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors". In: *Genome research* 28.5 (2018), pp. 654–665.

[15] Kin Chan et al. "An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers". In: *Nature genetics* 47.9 (2015), pp. 1067–1072.

[16] Shixiang Wang et al. "APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer". In: *Oncogene* 37.29 (2018), pp. 3924–3936.

[17] Wei Chong et al. "Association of clock-like mutational signature with immune checkpoint inhibitor outcome in patients with melanoma and NSCLC". In: *Molecular Therapy-Nucleic Acids* 23 (2021), pp. 89–100.

[18] Serena Nik-Zainal et al. "Mutational processes molding the genomes of 21 breast cancers". In: *Cell* 149.5 (2012), pp. 979–993.

[19] Valsamo Anagnostou et al. "Multimodal genomic features predict outcome of immune checkpoint blockade in non-small-cell lung cancer". In: *Nature Cancer* 1.1 (2020), pp. 99–111.

[20] Christopher D Steele et al. "Undifferentiated sarcomas develop through distinct evolutionary pathways". In: *Cancer Cell* 35.3 (2019), pp. 441–456.

[21] Christopher D Steele et al. "Signatures of copy number alterations in human cancer". In: *Nature* (2022), pp. 1–8.

[22] The Cancer Genome Atlas Research Network. "Comprehensive molecular profiling of lung adenocarcinoma". In: *Nature* 511 (2014), pp. 543–50.

[23] The Cancer Genome Atlas Research Network. "Comprehensive genomic characterization of squamous cell lung cancers". In: *Nature* 489.7417 (2012), p. 519.

[24] Andrea Degasperi et al. "Substitution mutational signatures in whole-genome–sequenced cancers in the UK population". In: *Science* 376.6591 (2022), abl9283.

[25] Jian-Min Chen, Claude Férec, and David N Cooper. "Patterns and mutational signatures of tandem base substitutions causing human inherited disease". In: *Human mutation* 34.8 (2013), pp. 1119–1130.

[26] Naiyer A Rizvi et al. "Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer". In: *Science* 348.6230 (2015), pp. 124–128.

[27] Yun Rose Li et al. "Mutational signatures in tumours induced by high and low energy radiation in Trp53 deficient mice". In: *Nature communications* 11.1 (2020), pp. 1–15.

[28] Mia Petljak et al. "Characterizing mutational signatures in human cancer cell lines reveals episodic APOBEC mutagenesis". In: *Cell* 176.6 (2019), pp. 1282–1294.

[29] Tamar Uziel et al. "Genomic organization of the ATM gene". In: *Genomics* 33.2 (1996), pp. 317–320.

[30] Kevin Litchfield et al. "Escape from nonsense-mediated decay associates with anti-tumor immunogenicity". In: *Nature communications* 11.1 (2020), pp. 1–11.

[31] H. C. Donker et al. "Decoding circulating tumor DNA to identify durable benefit from immunotherapy in lung cancer". In: *Lung Cancer* 170 (2022), pp. 52–57.

[32] Jan Norum and Carsten Nieder. "Tobacco smoking and cessation and PD-L1 inhibitors in non-small cell lung cancer (NSCLC): a review of the literature". In: *ESMO open* 3.6 (2018), e000406.

[33] Andrew J Vickers, Ben Van Calster, and Ewout W Steyerberg. "Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests". In: *bmj* 352 (2016).

[34] Andrew J Vickers, Ben van Calster, and Ewout W Steyerberg. "A simple, step-by-step guide to interpreting decision curve analysis". In: *Diagnostic and prognostic research* 3.1 (2019), pp. 1–8.

[35] Rajesh Talluri and Sanjay Shete. "Using the weighted area under the net benefit curve for decision curve analysis". In: *BMC medical informatics and decision making* 16.1 (2016), pp. 1–9.

[36] Peter Priestley et al. "Pan-cancer whole-genome analyses of metastatic solid tumours". In: *Nature* 575.7781 (2019), pp. 210–216.

[37] Daniel L. Cameron et al. "GRIDSS, PURPLE, LINX: Unscrambling the tumor genome via integrated analysis of structural variation and copy number". In: *bioRxiv* (2019). DOI: `10.1101/781013`. eprint: `https://www.biorxiv.org/content/early/2019/09/25/781013.full.pdf`. URL: `https://www.biorxiv.org/content/early/2019/09/25/781013`.

[38] Hao Zhao et al. "CrossMap: a versatile tool for coordinate conversion between genome assemblies". In: *Bioinformatics* 30.7 (2014), pp. 1006–1007.

[39] COSMIC. *COSMIC — Mutational Signatures*. 2022. URL: `https://cancer.sanger.ac.uk/signatures/` (visited on 05/17/2022).

[40] Erik N Bergstrom et al. "SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events". In: *BMC genomics* 20.1 (2019), pp. 1–12.

[41] Alison J Coffey et al. "The GENCODE exome: sequencing the complete human exome". In: *European journal of human genetics* 19.7 (2011), pp. 827–831.

[42] Min Zhao et al. "TSGene 2.0: an updated literature-based knowledge-base for tumor suppressor genes". In: *Nucleic acids research* 44.D1 (2016), pp. D1023–D1031.

[43] Zbyslaw Sondka et al. "The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers". In: *Nature Reviews Cancer* 18.11 (2018), pp. 696–705.

[44] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[45] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[46] Harry Zhang. "The optimality of naive Bayes". In: *Aa* 1.2 (2004), p. 3.

[47] Raymond Paternoster et al. "Using the correct statistical test for the equality of regression coefficients". In: *Criminology* 36.4 (1998), pp. 859–866.

[48] Eliezer M Van Allen et al. "Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine". In: *Nature medicine* 20.6 (2014), pp. 682–688.

[49] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (2014), pp. 2114–2120.

[50] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[51] Philip Ewels et al. "MultiQC: summarize analysis results for multiple tools and samples in a single report". In: *Bioinformatics* 32.19 (2016), pp. 3047–3048.

[52] Petr Danecek et al. "Twelve years of SAMtools and BCFtools". In: *Gigascience* 10.2 (2021), giab008.

[53] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python framework to work with high-throughput sequencing data". In: *bioinformatics* 31.2 (2015), pp. 166–169.

[54] Andrew J. Vickers and Elena B. Elkin. "Decision Curve Analysis: A Novel Method for Evaluating Prediction Models". In: *Medical Decision Making* 26.6 (2006), pp. 565–574.

[55] Emre Kocakavuk et al. "Radiotherapy is associated with a deletion signature that contributes to poor outcomes in patients with cancer". In: *Nature genetics* 53.7 (2021), pp. 1088–1096.