nature cancer

Article

https://doi.org/10.1038/s43018-024-00772-7

LORIS robustly predicts patient outcomes with immune checkpoint blockade therapy using common clinical, pathologic and genomic features

Received: 21 September 2023

Accepted: 24 April 2024

Published online: 3 June 2024

Check for updates

Tian-Gen Chang[1,7], Yingying Cao[1,7], Hannah J. Sfreddo[2,7], Saugato Rahman Dhruba[1], Se-Hoon Lee[3], Cristina Valero[2], Seong-Keun Yoo[4,5,6], Diego Chowell[4,5,6], Luc G. T. Morris[2] ✉ & Eytan Ruppin[1] ✉

Despite the revolutionary impact of immune checkpoint blockade (ICB) in cancer treatment, accurately predicting patient responses remains challenging. Here, we analyzed a large dataset of 2,881 ICB-treated and 841 non-ICB-treated patients across 18 solid tumor types, encompassing a wide range of clinical, pathologic and genomic features. We developed a clinical score called LORIS (logistic regression-based immunotherapy-response score) using a six-feature logistic regression model. LORIS outperforms previous signatures in predicting ICB response and identifying responsive patients even with low tumor mutational burden or programmed cell death 1 ligand 1 expression. LORIS consistently predicts patient objective response and short-term and long-term survival across most cancer types. Moreover, LORIS showcases a near-monotonic relationship with ICB response probability and patient survival, enabling precise patient stratification. As an accurate, interpretable method using a few readily measurable features, LORIS may help improve clinical decision-making in precision medicine to maximize patient benefit. LORIS is available as an online tool at https://loris.ccr.cancer.gov/.

Immune checkpoint blockade (ICB) has revolutionized our approach to cancer treatment. However, many patients do not respond to ICB therapy, creating a need to identify biomarkers to predict which patients may benefit from this treatment[1–3]. Although tumor mutational burden (TMB) has been recognized as a biomarker to predict ICB efficacy in solid tumors[4,5], current evidence fails to support the use of high TMB (with a US Food and Drug Administration (FDA)-approved threshold of 10 mutations per Mb) as a biomarker for response to ICB treatment universally, across all cancer types[6]. Other clinical, pathologic and genomic features reported to be associated with ICB response include programmed cell death 1 (PD-1) ligand 1 (PD-L1) expression in the tumor[7], microsatellite instability (MSI)[8–10], human leukocyte antigen class I (HLA-I) evolutionary divergence (HED)[11], loss-of-heterozygosity (LOH) status in HLA-I (ref. 12), fraction of copy number alteration (FCNA) or tumor aneuploidy[13,14], blood neutrophil–lymphocyte ratio (NLR)[15,16], blood albumin level[17], body mass index (BMI)[18], sex[19] and age[20]. Nonetheless, there remains an unmet need to identify factors for patient selection that are as readily measurable and provide more robust and accurate predictions for cancer ICB response and patient stratification than the approved TMB biomarker.

There have been a few attempts to integrate features from multiomics data into a single machine learning model to improve the predictive power of ICB response. For example, one study curated 55 unique biomarkers from the literature and used a tree-based ensemble

A full list of affiliations appears at the end of the paper. ✉e-mail: morrisl@mskcc.org; eytan.ruppin@nih.gov

model, identifying the 11 most predictive biomarkers for ICB response[21]. However, this approach relied on whole-exome sequencing (WES) and transcriptome sequencing, which are expensive approaches and not routinely measured in clinical settings. In another approach, a random forest model was developed using 16 genomic and clinical features to predict pan-cancer ICB response[22]. This model was tested only on data from the originating medical center, leaving open the challenge of additional testing on external and independent cohorts. Moreover, the 'black box' nature of these models has limited their interpretability, impeding their application in the clinic.

Because cancer drug response is a complex phenomenon, it is currently challenging to perfectly distinguish responders from nonresponders. Therefore, assessing the response probability of a patient to a particular therapy is of great value, potentially allowing clinicians to make more precise treatment decisions. For instance, in a patient with a high probability of ICB response, immunotherapy might be prioritized over another therapy; in a patient with a lower probability of ICB response, other therapeutic avenues might be prioritized. While TMB and PD-L1 expression are the two major FDA-approved biomarkers for ICB therapy, they do have limitations in accurately predicting response. For example, patients with low TMB may have a similar or even higher probability of responding to ICB therapy compared to those with high TMB[6,23]. Similarly, tumors across all PD-L1 expression levels may respond to ICB treatments[3,24]. Unfortunately, there has been little progress in developing a scoring system that can predict the patient-level ICB response probability in a monotonic manner, whereby higher scores reliably correlate with higher response probabilities across the entire score range.

Here, we developed and validated a transparent 'white box' computational model with a few clinically easy-to-measure features, which can help clinicians to determine the patient's probability of responding to ICB therapy. First, we curated and comprehensively analyzed a large collection of persons with different types of cancer, with more than 20 clinical, pathologic and genomic features measured. We then developed and tested 20 machine learning models using repeated cross-validation to identify the most predictive model for ICB response. Finally, we found that a clinical score, derived from a six-feature logistic regression model, had a superior and robust performance in predicting the objective response to ICB on both internal cross-validation and multiple independent datasets. Remarkably, this clinical score exhibited a monotonic relationship with both the ICB response probability, spanning from 0% to 100%, and the patient survival probability after treatment. Our findings suggest that this approach can be a powerful tool for predicting patient clinical outcomes in ICB therapy.

Results

Overview

We compiled a dataset of 2,881 participants with ICB treatment across 18 solid tumor types from multiple data sources (Fig. 1a, Table 1 and Extended Data Fig. 1a). All participants were treated with PD-1/PD-L1 inhibitors, cytotoxic T lymphocyte-associated protein 4 (CTLA-4) blockade or a combination of both immunotherapy agents (Fig. 1a and Table 1). To disentangle the predictive capacity of our model for ICB response from its prognostic significance within the broader context of cancer survival in the absence of ICB treatment, we also curated a cohort comprising 841 non-ICB-treated participants from 15 solid tumor types (Extended Data Fig. 1a,b).

The first data source was an MSK-IMPACT cohort, which included 1,479 participants treated for 16 cancer types at Memorial Sloan Kettering Cancer Center (MSK; Chowell et al. cohort[22]). The second data source was a cohort from South Korea, which included 198 participants with advanced non-small cell lung cancer (NSCLC) (Shim et al. cohort[25]). The third data source was an additional cohort from MSK, including 453 participants with 15 cancer types (MSK1 cohort) and 104 participants with either central nervous system (CNS) tumors or cancer of unknown primary (MSK2 cohort). The fourth data source was a pan-cancer study from the UCSD Moores Cancer Center, consisting of 35 participants across eight cancer types (Kato et al. cohort[26]), which matched those in the Chowell et al. cohort.

We also used data from the Vanguri et al. cohort[27], consisting of 246 participants with advanced NSCLC at MSK, the Stand Up To Cancer-Mark Foundation cohort, which included 309 participants with NSCLC (Ravi et al. cohort[28]), and a pan-cancer cohort of refractory metastatic tumors (META-PRISM, Pradat et al. cohort[29]) with 57 participants treated with ICB across 13 cancer types present in the Chowell et al. cohort. These very recently published studies were accessed and analyzed only after our model training and initial testing were completed and fixed.

To assess patient outcomes, three metrics were measured: objective response, progression-free survival (PFS) and overall survival (OS). Objective response was evaluated using the Response Evaluation Criteria in Solid Tumors (RECIST, version 1.1)[30] and classified as complete response (CR) or partial response (PR) for responders and stable disease (SD) or progressive disease (PD) for nonresponders. Among all ICB-treated participants, 825 (~29%) experienced an objective response while 2,056 (~71%) did not (Table 1). Across multiple cohorts, we evaluated more than 20 clinical, pathologic and genomic features (Methods). Among these features, eight were measured for most participants: sex, age, cancer type, ICB drug class, systemic therapy history, TMB, blood albumin level and blood NLR (Table 1). Here, 'systemic therapy history' was a binary variable indicating whether the participant received chemotherapy or targeted therapy before immunotherapy. Additionally, the PD-L1 tumor proportion score (TPS) was assessed in many NSCLC samples and a small portion of other cancer types.

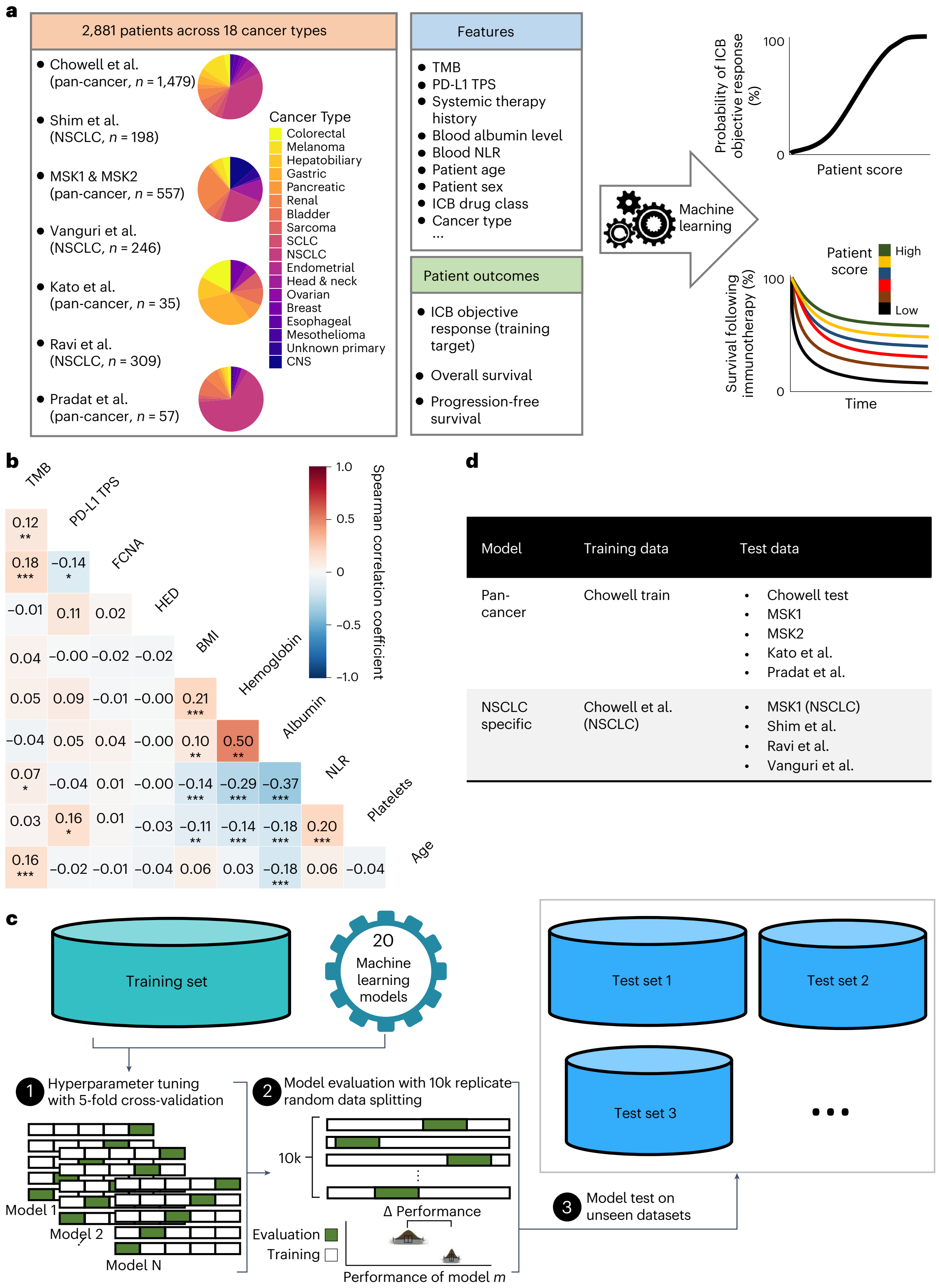
We first explored the correlation between features measured on a continuous scale at a pan-cancer level across all participants (Fig. 1b). TMB was positively correlated with FCNA ($r = 0.18$, adjusted $P < 0.001$) and age ($r = 0.16$, adjusted $P < 0.001$). PD-L1 TPS was positively correlated with blood platelets ($r = 0.16$, adjusted $P < 0.05$), which aligns with previous studies in ovarian cancer where platelets increased the expression of PD-L1 in tumors[31]. Interestingly, PD-L1 TPS was negatively correlated with FCNA ($r = -0.14$, adjusted $P < 0.05$). In addition, there was a strong positive correlation between blood hemoglobin and albumin level ($r = 0.50$, adjusted $P < 0.001$).

Next, we aimed to build a reliable ICB response predictor based on the measured features. To this end, we comprehensively built and evaluated response predictors using 20 different machine learning architectures. For each model, we first tuned the optimal hyperparameters using fivefold cross-validation on the training set; we then evaluated its performance using 2,000 repeats of fivefold cross-validation to ensure unbiased results (equating to 10,000 random training–validation splits in total). Finally, the selected models were further tested on multiple unseen test cohorts (Fig. 1c).

Our study included two types of models: pan-cancer and cancer type specific (Fig. 1d). Pan-cancer models were developed, trained, evaluated and compared using a subset of 964 participants from the Chowell et al. cohort who received immunotherapy between 2015 and 2017 (Chowell train). The unseen test cohorts included 515 participants from the Chowell et al. cohort who received immunotherapy in 2018 (Chowell test), as well as participants from the MSK1, MSK2, Kato et al. and Pradat et al. cohorts (Extended Data Fig. 1a). Cancer-type-specific models were developed, trained, evaluated and compared using the Chowell et al. cohort (only participants with NSCLC); unseen test cohorts included participants from the MSK1 (only participants with NSCLC), Shim et al., Ravi et al. and Vanguri et al. cohorts (Extended Data Fig. 1a). Overall, this approach allowed us to thoroughly evaluate the generalizability of the models under various scenarios.

A pan-cancer model to predict immunotherapy response

We first developed a pan-cancer logistic regression model to predict the objective response to ICB therapy using the eight features shared


a
2,881 patients across 18 cancer types
Chowell et al. (pan-cancer, n = 1,479)
Shim et al. (NSCLC, n = 198)
MSK1 & MSK2 (pan-cancer, n = 557)
Vanguri et al. (NSCLC, n = 246)
Kato et al. (pan-cancer, n = 35)
Ravi et al. (NSCLC, n = 309)
Pradat et al. (pan-cancer, n = 57)
Cancer Type
Colorectal
Melanoma
Hepatobiliary
Gastric
Pancreatic
Renal
Bladder
Sarcoma
SCLC
NSCLC
Endometrial
Head & neck
Ovarian
Breast
Esophageal
Mesothelioma
Unknown primary
CNS
Features
TMB
PD-L1 TPS
Systemic therapy history
Blood albumin level
Blood NLR
Patient age
Patient sex
ICB drug class
Cancer type
...
Patient outcomes
ICB objective response (training target)
Overall survival
Progression-free survival
Machine learning
Probability of ICB objective response (%)
100
0
Patient score
Survival following immunotherapy (%)
100
0
Time
Patient score
High
Low
b
TMB
PD-L1 TPS
FCNA
HED
BMI
Hemoglobin
Albumin
NLR
Platelets
Age
Spearman correlation coefficient
1.0
0.5
0
−0.5
−1.0
c
Training set
20 Machine learning models
1 Hyperparameter tuning with 5-fold cross-validation
Model 1
Model 2
Model N
2 Model evaluation with 10k replicate random data splitting
10k
Δ Performance
Evaluation
Training
Performance of model m
Test set 1
Test set 2
Test set 3
• • •
3 Model test on unseen datasets
d

	TMB	PD-L1 TPS	FCNA	HED	BMI	Hemoglobin	Albumin	NLR	Platelets
PD-L1 TPS	0.12 **								
FCNA	0.18 ***	−0.14 *							
HED	−0.01	0.11	0.02						
BMI	0.04	−0.00	−0.02	−0.02					
Hemoglobin	0.05	0.09	−0.01	−0.00	0.21 ***				
Albumin	−0.04	0.05	0.04	−0.00	0.10 **	0.50 **			
NLR	0.07 *	−0.04	0.01	−0.00	−0.14 ***	−0.29 ***	−0.37 ***		
Platelets	0.03	0.16 *	0.01	−0.03	−0.11 **	−0.14 ***	−0.18 ***	0.20 ***	
Age	0.16 ***	−0.02	−0.01	−0.04	0.06	0.03	−0.18 ***	0.06	−0.04

Model	Training data	Test data
Pan-cancer	Chowell train	• Chowell test • MSK1 • MSK2 • Kato et al. • Pradat et al.
NSCLC specific	Chowell et al. (NSCLC)	• MSK1 (NSCLC) • Shim et al. • Ravi et al. • Vanguri et al.

Fig. 1 | Overview of the study. a, Description of the study aims and data used. The study aimed to develop and validate machine learning models to predict patient objective response probability and survival benefit following immunotherapy. **b**, Correlation among features measured on a continuous scale at the pan-cancer level (*n* = 2,881 participants). *P* values were determined by Spearman's rank test, adjusted by Bonferroni correction. *, adjusted $P < 0.05$; **, adjusted $P < 0.01$; ***, adjusted $P < 0.001$. **c**, Schematic representation of the training, validation and independent testing procedures used to develop and evaluate the predictive models. For each machine learning architecture, the hyperparameter was tuned with fivefold cross-validation. After determination of the hyperparameters, the models were evaluated using various performance metrics with 2,000 repeats of fivefold cross-validation. Lastly, the selected models were tested on multiple unseen test cohorts to assess their generalizability. **d**, The two types of models built, that is, the pan-cancer and NSCLC-specific models, and the corresponding training and test data used.

Table 1 | Characteristics of participants with ICB treatment in the study

Characteristic	Total participants	Chowell et al. cohort[22]	Shim et al. cohort[25]	MSK1 cohort	MSK2 cohort	Vanguri et al. cohort[27]	Kato et al. cohort[26]	Ravi et al. cohort[28]	Pradat et al. cohort[29]
Sex, *n* (%)									
Female	1,280 (44.4)	668 (45.2)	58 (29.3)	172 (38.0)	42 (40.4)	134 (54.5)	17 (48.6)	165 (53.3)	24 (42.1)
Male	1,601 (55.6)	811 (54.8)	140 (70.7)	281 (62.0)	62 (59.6)	112 (45.5)	18 (51.4)	144 (46.6)	33 (57.9)
Age, median, years (IQR)	63 (55–71)	64 (55–71)	62 (55–69)	63 (53–73)	53 (48–63)	68 (61–73)	62 (51–72)	64 (57–71)	66 (54–69)
Cancer type, *n* (%)									
NSCLC	1,456 (50.5)	538 (36.4)	198 (100)	128 (28.3)	–	246 (100)	–	309 (100)	37 (64.9)
Renal	232 (8.1)	91 (6.2)	–	137 (30.2)	–	–	–	–	4 (7.0)
Melanoma	217 (7.5)	186 (12.6)	–	30 (6.6)	–	–	–	–	1 (1.8)
Head and neck	132 (4.6)	69 (4.7)	–	61 (13.5)	–	–	2 (5.7)	–	–
Bladder	119 (4.1)	82 (5.5)	–	29 (6.4)	–	–	3 (8.6)	–	5 (8.8)
Sarcoma	88 (3.1)	67 (4.5)	–	17 (3.8)	–	–	3 (8.6)	–	1 (1.8)
Gastric	82 (2.8)	64 (4.3)	–	7 (1.5)	–	–	11 (31.4)	–	–
CNS	75 (2.6)	–	–	–	75 (72.1)	–	–	–	–
Colorectal	75 (2.6)	46 (3.1)	–	22 (4.9)	–	–	6 (17.1)	–	1 (1.8)
Endometrial	71 (2.5)	65 (4.4)	–	4 (0.9)	–	–	–	–	2 (3.5)
Hepatobiliary	62 (2.2)	52 (3.5)	–	5 (1.1)	–	–	4 (11.4)	–	1 (1.8)
CLC	55 (1.9)	50 (3.4)	–	4 (0.9)	–	–	–	–	1 (1.8)
Esophageal	50 (1.7)	44 (3.0)	–	5 (1.1)	–	–	–	–	1 (1.8)
Pancreatic	40 (1.4)	35 (2.4)	–	1 (0.2)	–	–	3 (8.6)	–	1 (1.8)
Mesothelioma	36 (1.2)	34 (2.3)	–	1(0.2)	–	–	–	–	1 (1.8)
Ovarian	31 (1.1)	31 (2.1)	–	–	–	–	–	–	–
Breast	31 (1.1)	25 (1.7)	–	2 (0.4)	–	–	3 (8.6)	–	1 (1.8)
Unknown primary	29 (1.0)	–	–	–	29 (27.9)	–	–	–	–
Drug class, *n* (%)									
PD-1/PD-L1	2,447 (86.0)	1,221 (82.6)	198 (100)	390 (86.1)	102 (98.1)	234 (95.1)	–	245 (79.3)	57 (100)
CTLA-4	7 (0.2)	5 (0.3)	–	2 (0.4)	–	–	–	–	–
Combo	392 (13.8)	253 (17.1)	–	61 (13.5)	2 (1.9)	12 (4.9)	–	64 (20.7)	–
Systemic therapy history, *n* (%)									
No	814 (28.3)	463 (31.3)	14 (7.1)	107 (23.6)	24 (24)	78 (31.7)	–	123 (39.8)	5 (8.8)
Yes	2,063 (71.7)	1,016 (68.7)	184 (92.9)	346 (76.4)	76 (76)	168 (68.3)	35 (100)	186 (60.2)	52 (91.2)
TMB, median, mutations per Mb (IQR)	5.9 (3.0–10.8)	5.3 (2.8–10.8)	7.3 (3.7–12.0)	5.3 (3.3–8.9)	3.9 (2.5–5.3)	7.9 (4.4–12.3)	7 (5–11)	7.4 (3.4–12.7)	7.1 (1.8–12.2)
Albumin, median, g dl^{-1} (IQR)	3.9 (3.6–4.2)	3.9 (3.6–4.1)	4.1 (3.8–4.4)	3.9 (3.6–4.2)	4.1 (3.8–4.3)	3.8 (3.4–4.1)	–	–	3.9 (3.6–4.1)
NLR, median, (IQR)	4.3 (2.7–7.0)	4.4 (2.8–7.2)	3.1 (1.9–4.8)	4.1 (2.7–7.1)	4.1 (2.5–6.8)	4.9 (3.3–7.9)	–	–	4.5 (2.9–6.3)
PD-L1 TPS, median, % (IQR)	5 (0–66)	0 (0–60)	50 (1–72.5)	1 (0–50)	0 (0–50)	5 (0–60)	–	25 (0–75)	–
ICB response, *n* (%)									
Responder	825 (28.6)	409 (27.7)	61 (30.8)	116 (25.6)	14 (13.5)	61 (24.8)	5 (14.3)	121 (39.2)	19 (33.3)
Nonresponder	2,056 (71.4)	1,070 (72.3)	137 (69.2)	337 (74.4)	90 (86.5)	185 (75.2)	30 (85.7)	188 (60.8)	38 (66.7)

among all participants. We performed fivefold cross-validation to identify the optimal hyperparameters. Feature importance analysis showed that a participant's sex and ICB drug class information had little impact on the prediction (Extended Data Fig. 1c). After removing these two features, the best model found was a six-feature logistic LASSO (least absolute shrinkage and selection operator) regression model (LLR6), which included the following features in decreasing order of importance: TMB, systemic therapy history, blood albumin, blood NLR, age and cancer type (Extended Data Fig. 1d).

To double-check and assess the possible added value of the other features, we also developed and tuned a logistic regression model using all 16 features. However, the 16-feature model showed no improvement in performance over LLR6 on the cross-validation sets (Supplementary Table 1). In addition, we tested a five-variable logistic regression model without using TMB, LR5 (noTMB), but it performed worse than LLR6 (Supplementary Table 1). Overall, our analysis suggested that the six selected features captured the most essential information for predicting ICB response in participants with different types of cancer.

We also compared the performance of LLR6 with an established method, referred to as RF16 (Chowell et al.) hereafter, which is a 16-feature random forest model reported recently[22]. As a result, LLR6 outperformed RF16 (Chowell et al.) by having significantly higher values in five of seven different metrics on the cross-validation sets (Extended Data Fig. 2a). Notably, LLR6 exhibited a close-to-zero performance difference between training and cross-validation, much smaller than that of RF16 (Chowell et al.) (Extended Data Fig. 2b), which suggests that LLR6 is less prone to overfitting than RF16 (Chowell et al.). Indeed, while RF16 (Chowell et al.) exhibited significantly higher area under the receiver operating characteristic curve (AUC) and area under the precision–recall curve (AUPRC) values than LLR6 on the training data, it experienced a substantial drop in performance on the unseen test data, ultimately resulting in even poorer performance than LLR6 (Extended Data Fig. 2c). Interestingly, while the scores calculated from LLR6 and RF16 (Chowell et al.) were highly correlated, LLR6 scores exhibited a more uniform distribution across the range of 0 to 1. In contrast, scores generated from RF16 (Chowell et al.) tended to cluster within a narrower range between 0 and 0.6 (Extended Data Fig. 2d).

To further test if there were better machine learning architectures for predicting ICB response using all 16 features, we experimented with 15 additional machine learning models, such as decision trees, Gaussian processes, support vector machine, XGBoost and deep neural networks. However, none of these models outperformed LLR6. While some complex models, such as XGBoost and a two-layer multilayer perceptron network, showed comparable performance to LLR6 (Supplementary Table 1), they exhibited much larger discrepancies in performance between the training and validation data (Supplementary Table 2), indicating a high risk of overfitting the data. We also compared other clinical and computational characteristics of LLR6 with other models. In short, LLR6 was the best model that simultaneously possessed the desirable properties of (1) superior performance and being less prone to overfitting; (2) use of only a few clinically measurable features; (3) high transparency and interpretability; and (4) short model training time, among others (Supplementary Table 3).

To directly assess the generalizability of LLR6, we applied it to five unseen datasets. We referred to the output calculated using LLR6 as the logistic regression-based immunotherapy-response score (LORIS) (Methods). As a baseline, we tested the FDA-approved TMB biomarker. Additionally, while we were unable to directly evaluate RF16 (Chowell et al.) on these external datasets because of the absence of many input features, we successfully constructed a six-feature random forest model (RF6) and optimized its hyperparameters using the same protocol as for the development of RF16 (Chowell et al.)[22].

LLR6 consistently outperformed RF6 and the TMB biomarker across all datasets, even for cancer types not seen in the training data such as CNS tumors and cancer of unknown primary in the MSK2 cohort. Specifically, LLR6 achieved 1–39% and 15–68% higher AUCs than RF6 and the TMB biomarker, respectively (Fig. 2a). Simultaneously, LLR6 consistently outperformed RF6 and the TMB biomarker by predicting significantly higher LORIS for responders compared to nonresponders on all datasets (Fig. 2b). In addition, LLR6 showed superior AUPRCs on most datasets (Fig. 2c).

To binarize the values of LORIS and RF6 scores, we used cutoffs of 0.5 and 0.27, respectively, which maximized Youden's index, defined as 'sensitivity + specificity − 1', of the models on the training data, respectively. Regarding TMB, the FDA-approved cutoff of 10 mutations per Mb was used. Using binarized scores, LLR6 predicted an odds ratio of 1.4–4.1 for ICB objective response between high-LORIS and low-LORIS participants, which was higher than for RF6 (1.1–3.5) and the TMB biomarker (0.8–2.6) (Fig. 2c).

LORIS identifies low-TMB responders to immunotherapy

We further studied whether LORIS could predict patient survival outcomes following immunotherapy. Our pan-cancer Kaplan–Meier analysis revealed that participants with low LORIS (binned at 0.5) had significantly worse survival compared to those with high scores (OS: hazard ratio (HR) = 3.2, 95% confidence interval (CI) = 2.6–3.9, $P = 2 \times 10^{-28}$; Fig. 3a; PFS: HR = 2.6, 95% CI = 2.2–3.0, $P = 2 \times 10^{-33}$; Extended Data Fig. 3a). In contrast, using TMB (binned at 10 mutations per Mb) to stratify participants resulted in moderate power (OS: HR = 1.3, 95% CI = 1.1–1.6, $P = 0.01$; Fig. 3a; PFS: HR = 1.5, 95% CI = 1.2–1.8, $P = 8 \times 10^{-6}$; Extended Data Fig. 3a). Notably, LORIS identified a substantial proportion of low-TMB participants who could benefit from immunotherapy at a similar level to high-TMB participants (Fig. 3a and Extended Data Fig. 3a). Similar results were obtained when we used the 50th percentile in each cancer type as the optimal cutoff for LORIS binning and the highest 20th percentile for TMB binning (Fig. 3b and Extended Data Fig. 3b). Note that the highest 20th percentile was used as it was the optimal threshold for TMB binning proposed in a previous study[5]. Indeed, when using the 50th percentile for TMB binning, akin to the approach used for LORIS, a more pronounced trend emerged. Specifically, participants with high LORIS scores, regardless of whether they exhibited low or high TMB levels, tended to derive similar benefits from immunotherapy. Conversely, participants with low LORIS scores had a limited potential to benefit from immunotherapy (Extended Data Fig. 4a,b).

To test the predictive power of LLR6 in individual cancer types, we calculated HRs for LORIS and TMB for each cancer type using multivariate Cox proportional hazards regression that accounted for age, ICB drug class and year of ICB start. Consequently, higher LORIS predicted better OS (HR < 1) for all except one individual cancer type (binned at the 50th percentile; Fig. 3c), which was not true for the TMB biomarker (binned at the highest 20th percentile; Fig. 3d). Similar results were also observed for PFS (Extended Data Fig. 3c,d). Consistently, Kaplan–Meier analyses show that survival following immunotherapy was worse in low-LORIS participants for all 18 individual cancer types (Extended Data Fig. 5).

We also examined whether higher LORIS could predict better short-term and long-term patient survival, as both metrics are clinically important on their own. We compared the survival probability between participants with high versus low LORIS at various time points after immunotherapy, including half a year, 1 year, 2 years, 3 years, 4 years and 5 years. Notably, higher LORIS predicted significantly better OS for all time points (difference in survival probability between high-LORIS and low-LORIS participants: 0.21–0.33; Wilcoxon test P values: $3 \times 10^{-5} - 3 \times 10^{-3}$; Fig. 3e). In contrast, TMB did not consistently predict better OS for all time points (Fig. 3f). We also observed similar results for PFS (Extended Data Fig. 3e,f).

LORIS provides monotonic response and survival prediction

Next, we investigated the relationship between a participant's LORIS and their outcomes. Notably, we uncovered a unique characteristic of the LORIS signature. Specifically, as the LORIS increased, there was a consistent rise in the probability of objective response for participants, ranging from 0% to 100% (Fig. 4a). This distinctive attribute would allow clinicians to easily estimate the likelihood of ICB response of a person with cancer by assessing the six input features. In particular, LORIS enabled identification of the top 10% of participants who were highly likely to respond to ICB therapy (with a response probability exceeding 50%) while excluding the bottom ~10% of participants who were unlikely to respond (with a response probability below 10%). In contrast, the stratification power of TMB fell short. Only the top 6% of participants with the highest TMB exhibited a response probability exceeding 50%, while the lowest TMB scores proved ineffective in excluding nonresponsive participants altogether (Fig. 4b).

We further found that a higher LORIS consistently predicted better OS for participants across different percentiles. We were able to group patient survival into as many as six categories based on LORIS within each cancer type, that is, 0–10%, <10–20%, <20–50%, <50–80%, <80–90%

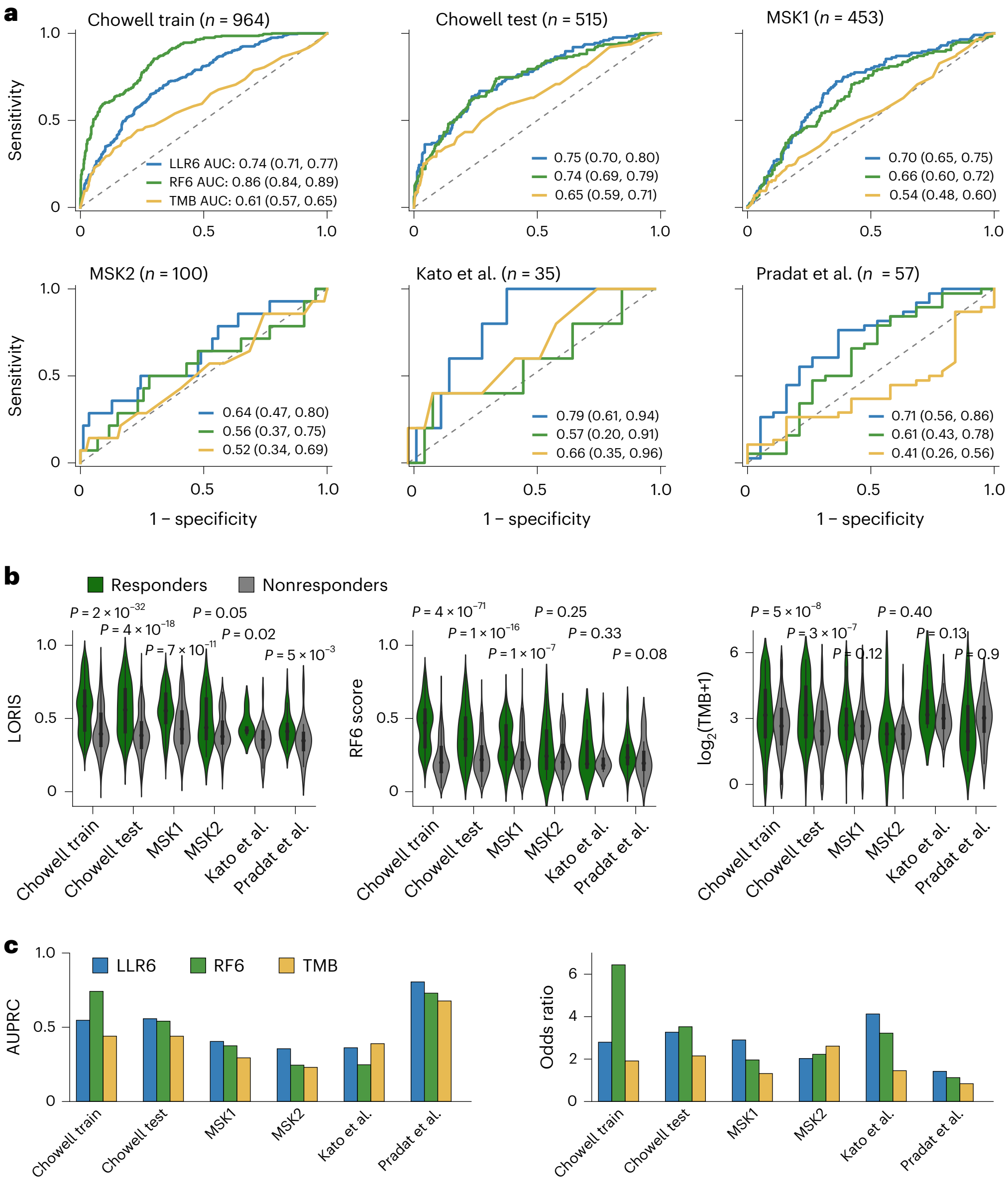


Fig. 2 | Robust prediction of pan-cancer objective response to immunotherapy by a six-variable LLR model. a, Receiver operating characteristic curves and corresponding AUCs with 95% CIs of LLR6 (blue curves), RF6 (green curves) and the TMB biomarker (yellow curves) on the training set and across multiple unseen test sets. In the figure, *n* represents the number of participants. The dashed lines represent random performance, serving as a baseline with an AUC of 0.5. This indicates the performance expected from a classifier making random guesses. **b**, Distribution of LORIS, RF6 score and TMB alone in responders and nonresponders on the training set and across multiple unseen test sets. *P* values were determined by a two-tailed Mann–Whitney *U* test. Box boundaries represent the first and third quartiles; the central line marks the median. Whiskers extend to the furthest nonoutlier points within 1.5 times the interquartile range. Outliers are shown as points beyond the whiskers. **c**, AUPRCs and odds ratios of the ICB objective response of LLR6 (blue bars), RF6 (green bars) and the TMB biomarker (yellow bars) on the training set and across multiple unseen test sets. The number of participants in different cohorts is displayed in **a**.

and <90–100%. Notably, the HR between the lowest-percentile (0–10%) and highest-percentile (90–100%) groups was as high as 7.8 (95% CI = 4.9–12.4, $P = 1 \times 10^{-18}$; Fig. 4c). However, we did not observe this monotonic relationship with survival for the TMB biomarker (Fig. 4d). Similar results were observed for PFS (Extended Data Fig. 4c,d).

LORIS has enhanced predictive power over prognosis

As biomarkers may have prognostic value, predictive value or both[32], we next sought to explore the degree to which LORIS could prognosticate patient outcome outside of the context of ICB therapy. To explore this, LORIS scores were calculated for a cohort of participants with cancer from 15 solid tumor types ($n = 841$) that were treated with standard therapies (non-ICB cohort) at MSK. While LORIS had moderate prognostic value of patient survival in the non-ICB setting (AUCs for 0.5-year to 3-year OS: 0.60–0.61), the AUCs were significantly lower than those observed for the ICB therapy group (AUCs: 0.73–0.83, DeLong's test $P \leq 1 \times 10^{-7}$; Fig. 5a). Additionally, the correlation between higher LORIS scores and improved survival in non-ICB-treated participants lacked a clear monotonic trend, with a much smaller risk difference between highest-scored and lowest-scored participants (HR = 1.3, 95% CI = 0.8–2.1, $P = 0.13$; Fig. 5b). In individual cancer types, LORIS also exhibited a reduced risk estimation capacity (Fig. 5c) and less statistical power in distinguishing short-term and/or long-term patient survival (Fig. 5d) within the non-ICB cohort. In addition, we conducted comparative analysis using the LR5 (noTMB) model, excluding the TMB component. The LR5 (noTMB) model continued to demonstrate

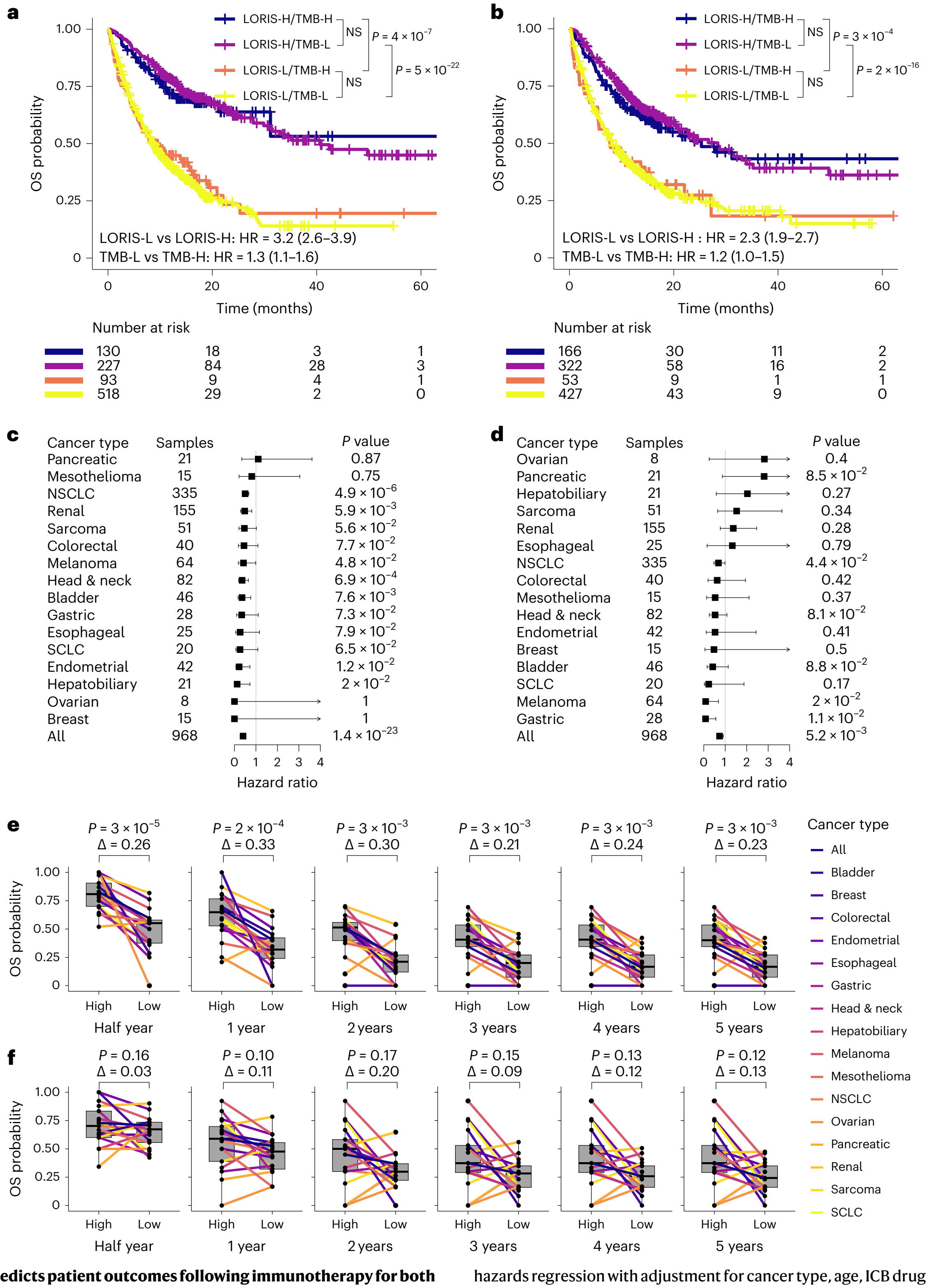


Fig. 3 | LORIS predicts patient outcomes following immunotherapy for both pan-cancer and individual cancer types. a, Kaplan–Meier analysis of OS. TMB is binned at 10 mutations per Mb and LORIS is binned at 0.5. HRs with 95% CIs are shown. *P* values were determined by univariable Cox proportional hazards regression. H, high; L, low. In the risk table, the numbers represent the numbers of participants. **b**, Same as **a** but TMB is binned at the highest 20th percentile and LORIS is binned at the 50th percentile for each cancer type. HRs with 95% CIs are shown. *P* values were determined by univariable Cox proportional hazards regression. **c**,**d**, Forest plot of HRs of OS within each cancer type using LORIS (binned at the 50th percentile) (**c**) or TMB (binned at the highest 20th percentile) (**d**). *P* values were determined by multivariable Cox proportional hazards regression with adjustment for cancer type, age, ICB drug class and year of ICB start. Squares positioned at midpoints symbolize point estimates of HRs and the accompanying bars indicate 95% CIs. **e**,**f**, Comparison of half-year, 1-year, 2-year, 3-year, 4-year and 5-year OS stratified by cancer type for high versus low LORIS (binned at the 50th percentile) (**e**) and high versus low TMB (binned at the highest 20th percentile) (**f**). Median survival probability differences (Δ) are displayed. *P* values were determined by two-tailed paired Wilcoxon rank sum test. Box boundaries represent the first and third quartiles; the central line marks the median. Whiskers extend to the furthest nonoutlier points within 1.5 times the interquartile range. Data are from the combined Chowell test and MSK1 sets ($n = 968$ participants).

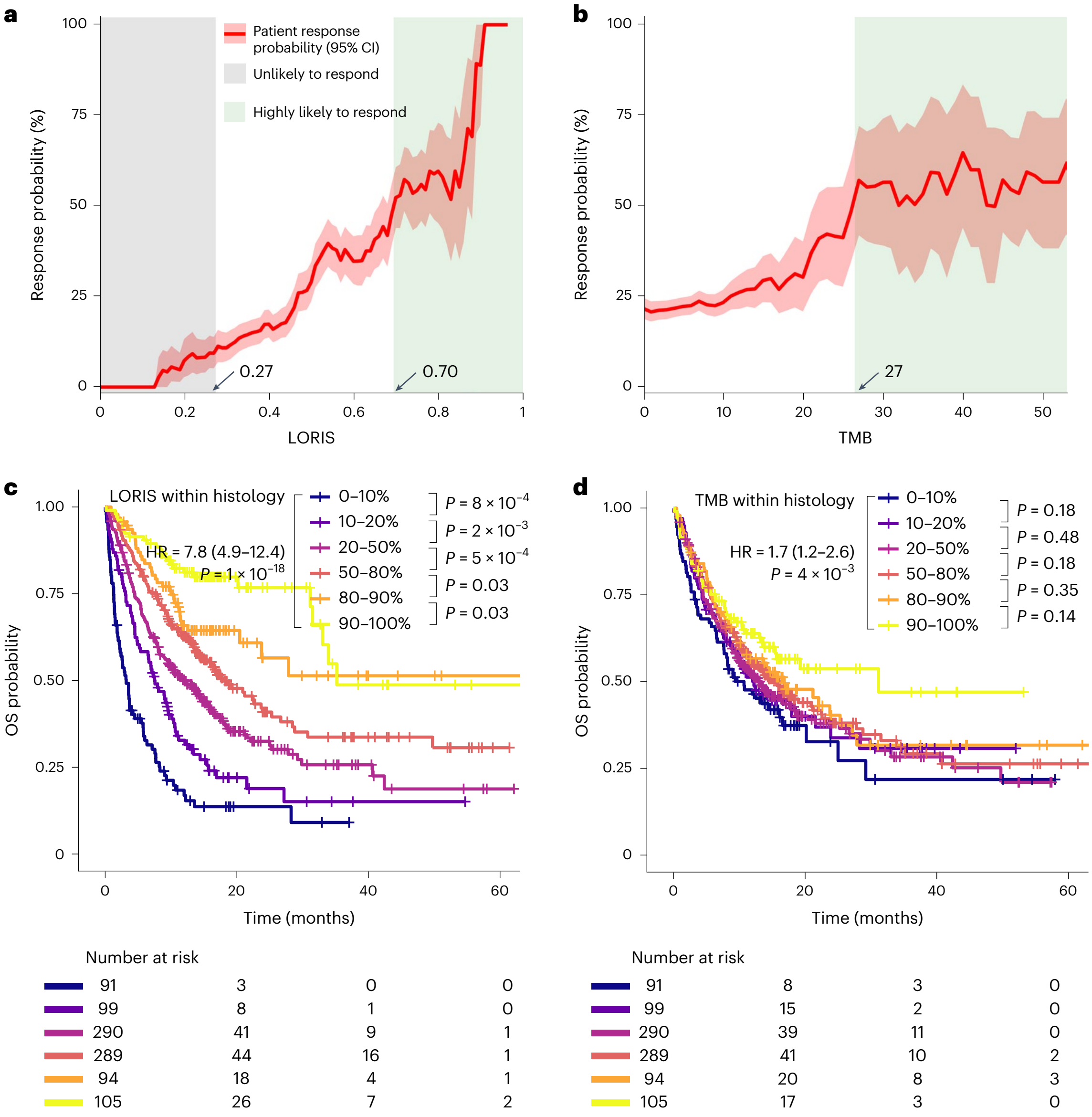


Fig. 4 | Monotonic relationship between LORIS and patient objective response probability and survival following immunotherapy. **a**,**b**, Relationship between LORIS (**a**) or TMB (**b**) and ICB objective response probability. The average patient response probabilities with 95% CIs are shown using 1,000 bootstrap replicates. The gray region represents participants with an unlikely response to immunotherapy (with a response probability below 10%), while the green regions represent participants with a likely response (with a response probability exceeding 50%). The arrows indicate the LORIS and TMB threshold values. **c**,**d**, Kaplan–Meier analysis of LORIS (**c**) or TMB (**d**) binned at the different percentiles in each cancer type. *P* values next to the legend indicate pairwise single-tail comparisons testing against the hypothesis that 'higher-scored participants do not have better survival than lower-scored participants' with univariable Cox proportional hazards regression. HRs with 95% CIs are shown for the lowest-percentile (0–10%) and the highest-percentile (90–100%) groups with univariable Cox proportional hazards regression. Data are from the combined Chowell test and MSK1 sets (*n* = 968 participants).

superior predictive power over its prognostic ability (Extended Data Fig. 4e). Taken together, these results showed that LORIS has both prognostic and ICB treatment predictive value but its predictive value appears much stronger than, and not attributable to, its value as a general prognostic marker in persons with cancer.

Enhancing lung cancer immunotherapy predictions with LORIS

Our study demonstrated the superior capability of our pan-cancer model in predicting ICB response. This success led us to probe a subsequent question: Could the approach be extended to develop cancer-type-specific models? To this end, we tested the potential of using logistic LASSO regression (LLR) to create a specific model for NSCLC, the cancer type with the largest sample size in our dataset.

We constructed, trained and assessed NSCLC-specific models using a similar protocol to our pan-cancer study, albeit with two minor adjustments. First, we harnessed the entire Chowell et al. cohort as our training data to ensure an adequate number of samples for model training. Secondly, we replaced the cancer type feature in the pan-cancer LLR6 model with PD-L1 TPS, as the former was redundant for a single-cancer study and the latter is a key biomarker routinely measured in persons with NSCLC. Consequently, a total of 324 participants with NSCLC in the training dataset were evaluated with complete measurement of the six input features.

As a result, the NSCLC-specific LLR6 model was one of the best models with 2,000 repeats of fivefold cross-validation compared to the 19 other models (Supplementary Table 4). More importantly, despite the limited data size, it maintained the desirable property of a near-zero

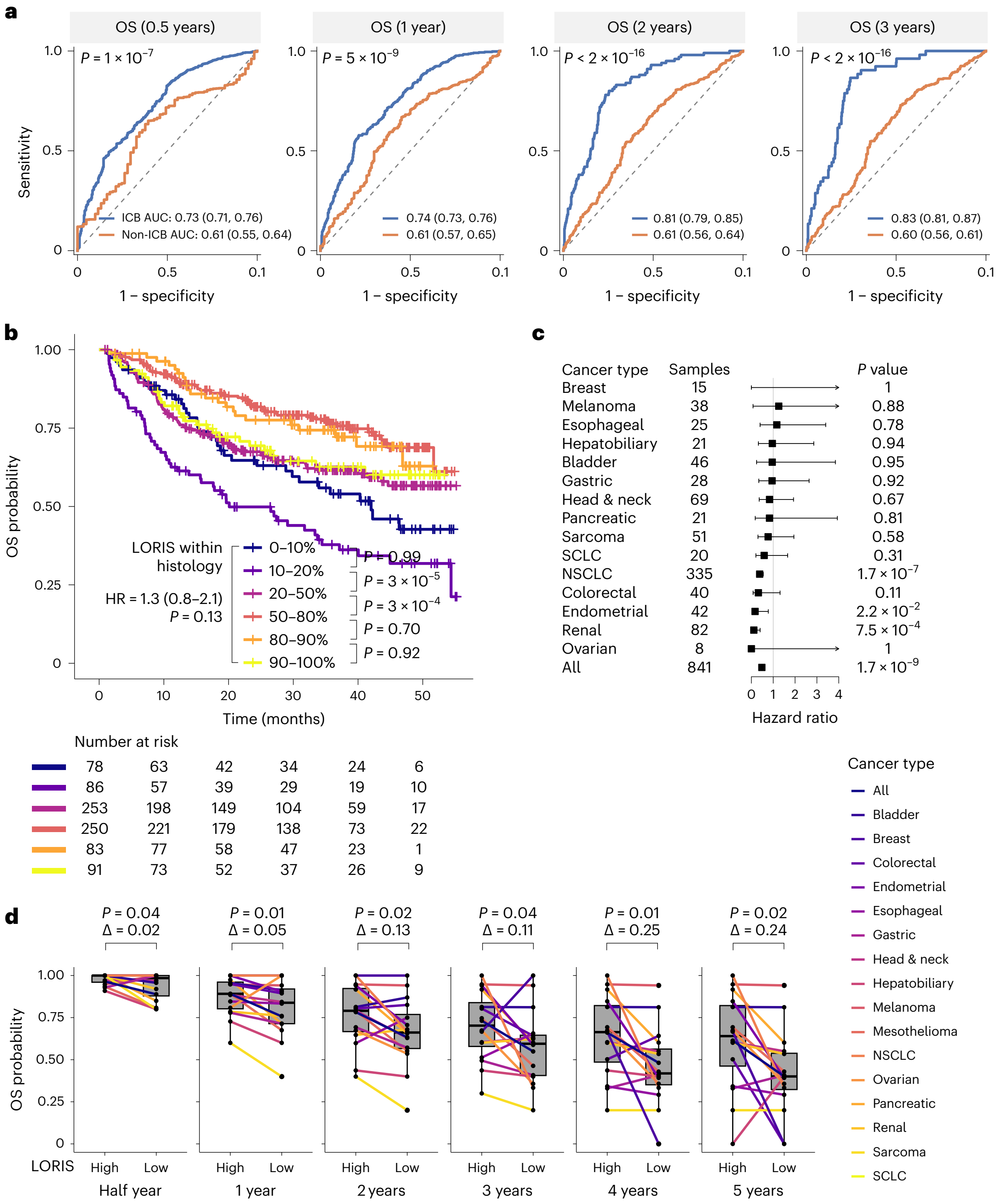


Fig. 5 | LORIS exhibits enhanced predictive efficacy for immunotherapy with respect to its prognostic value in the context of non-ICB treatments. **a**, Receiver operating characteristic curves and corresponding AUCs with 95% CIs of LORIS on 0.5-year, 1-year, 2-year and 3-year OS of participants treated with ICB (blue curves) or non-ICB (orange curves) therapies. *P* values were determined by two-tailed DeLong's test (non-ICB, *n* = 841 participants; ICB, *n* = 968 participants). The dashed lines represent random performance, serving as a baseline with an AUC of 0.5. This indicates the performance expected from a classifier making random guesses. **b**, Kaplan–Meier analysis of LORIS binned at the different percentiles in each cancer type for the non-ICB cohort. *P* values were determined by univariable Cox proportional hazards regression (single tail). HRs with 95% CIs are shown for the lowest-percentile (0–10%) and the highest-percentile (90–100%) groups (*n* = 841 participants). **c**, Forest plot of HRs of OS within each cancer type using LORIS (binned at the 50th percentile) for the non-ICB cohort. *P* values were determined by multivariable Cox proportional hazards regression with adjustment for cancer type and age. Squares positioned at midpoints symbolize the point estimates of HRs and the accompanying bars indicate the 95% CIs. **d**, Comparison of half-year, 1-year, 2-year, 3-year, 4-year and 5-year OS stratified by cancer type for high versus low LORIS (binned at the 50th percentile) for the non-ICB cohort. Median survival probability differences (Δ) are displayed. *P* values were determined by two-tailed paired Wilcoxon rank sum test. Box boundaries represent the first and third quartiles; the central line marks the median. Whiskers extend to the furthest nonoutlier points within 1.5 times the interquartile range. ICB data are from the combined Chowell test and MSK1 sets (*n* = 968 participants). Non-ICB data are from the MSK non-ICB cohort (*n* = 841 participants; Extended Data Fig. 1).

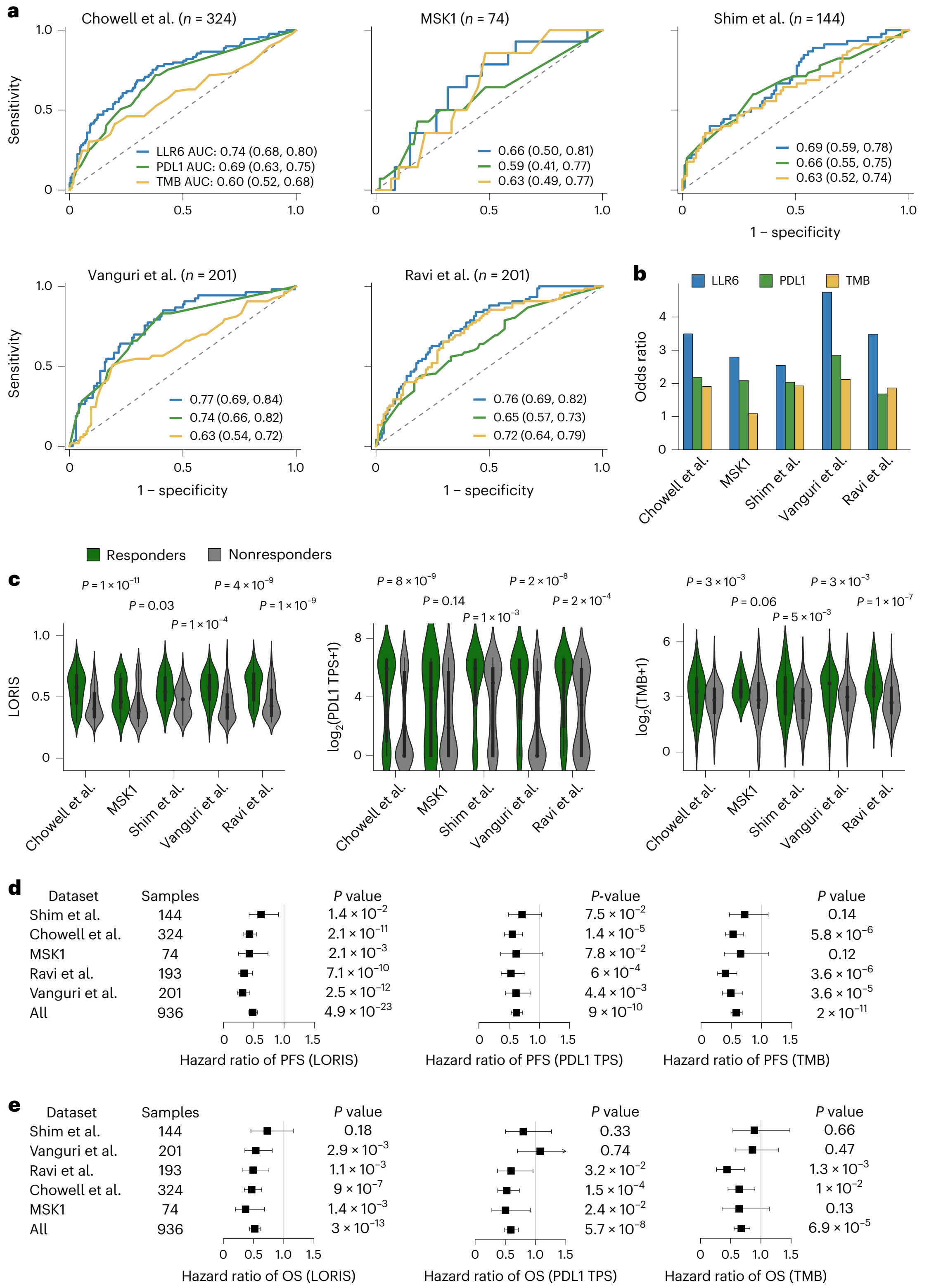


Fig. 6 | Robust prediction of response to immunotherapy in NSCLC with LLR. **a**, Receiver operating characteristic curves and corresponding AUCs with 95% CIs of LLR6 (blue curves), PD-L1 TPS (green curves) and TMB (yellow curves) on the training set and across multiple unseen test sets. In the figure, *n* represents the number of participants. The dashed lines represent random performance, serving as a baseline with an AUC of 0.5. This indicates the performance expected from a classifier making random guesses. **b**, Odds ratio of ICB objective response of LLR6 (blue bars), PD-L1 TPS (green bars) and TMB (yellow bars) on the training set and across multiple unseen test sets. **c**, Distribution of LORIS, PD-L1 TPS and TMB in responders and nonresponders on the training set and across multiple unseen test sets. *P* values were determined by a two-tailed Mann–Whitney *U* test. Box boundaries represent the first and third quartiles; the central line marks the median. Whiskers extend to the furthest nonoutlier points within 1.5 times the interquartile range. Outliers are shown as points beyond the whiskers. **d**,**e**, Forest plots of HRs of PFS (**d**) and OS (**e**) within each dataset using LORIS (binned at 0.44, which maximized Youden's index on the training data), PD-L1 TPS (binned at 50%) or TMB (binned at 10 mutations per Mb). *P* values were determined by multivariable Cox proportional hazards regression with adjustment for sex, age and ICB drug class. Squares positioned at midpoints symbolize the point estimates of HRs and the accompanying bars indicate the 95% CIs. The number of participants in different cohorts is displayed in **a**.

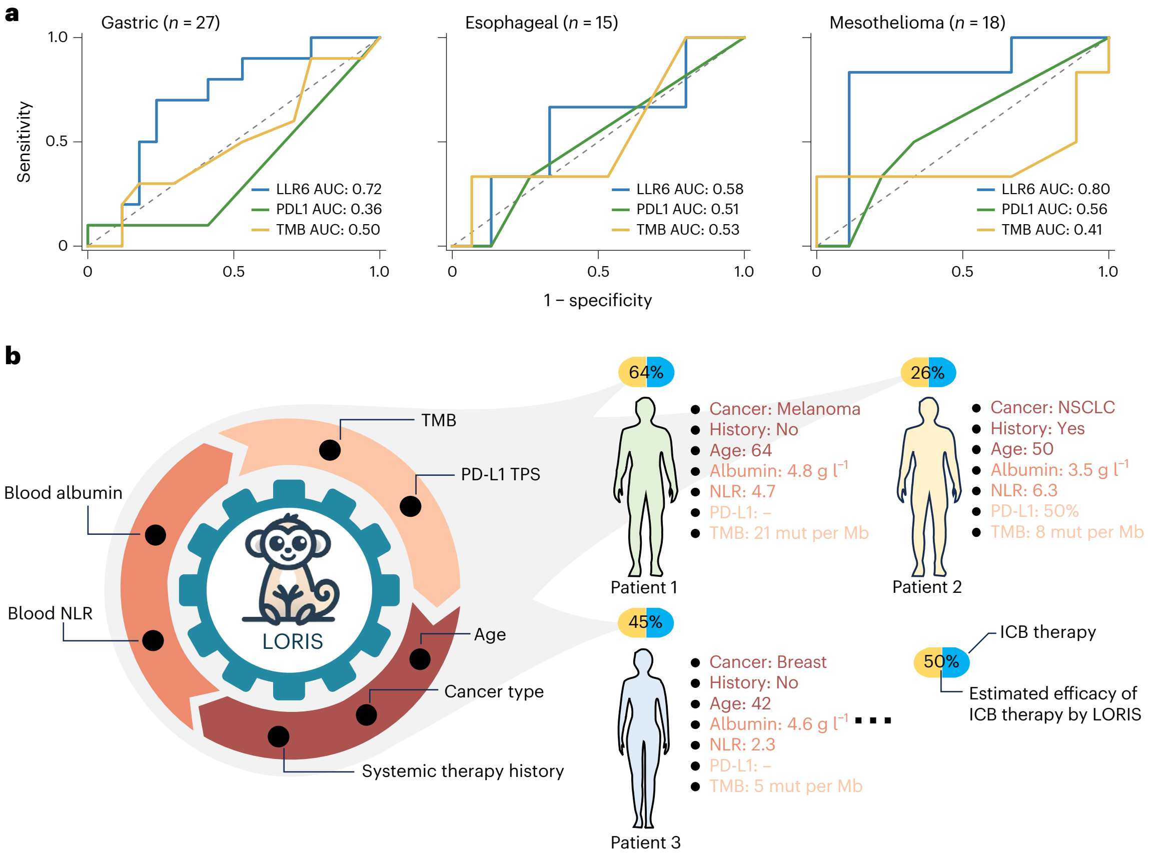


Fig. 7 | LORIS facilitates more precise ICB response prediction. a, Receiver operating characteristic curves and corresponding AUCs of the NSCLC-specific LLR6 model (blue curves), the PD-L1 TPS biomarker (green curves) and the TMB biomarker (yellow curves) on gastric cancer, esophageal cancer and mesothelioma. In the figure, *n* represents the number of participants. The dashed lines represent random performance, serving as a baseline with an AUC of 0.5. This indicates the performance expected from a classifier making random guesses. **b**, A summary of this study. LORIS, a clinical score derived from this study, estimates ICB response probabilities using LLR that identifies and integrates a few key features from three categories: tumor molecular data, blood measurements and patient clinical information. LORIS provides precise, patient-specific predictions of ICB therapy efficacy.

performance discrepancy between the training and cross-validation data (Supplementary Table 5). This suggests a minimized risk of overfitting. Indeed, NSCLC-specific LLR6 consistently outperformed both the TMB and the PD-L1 TPS biomarkers on all five external datasets, achieving 4–17% and 5–23% higher AUCs, respectively (Fig. 6a).

Moreover, the NSCLC-specific LORIS predicted an odds ratio of 2.5–4.7 for ICB response between high-LORIS and low-LORIS participants, which is much higher than predicted for the TMB (1.1–2.1) and PD-L1 TPS (1.7–2.9) biomarkers (Fig. 6b). In addition, responders consistently had significantly higher LORIS than nonresponders across all datasets (Fig. 6c). Lastly, higher LORIS consistently predicted a lower risk (HRs < 1) for both PFS and OS on all datasets, after adjusting for sex, age and ICB drug class (Fig. 6d,e).

To assess the additional value of constructing cancer-type-specific models, we conducted a comparative analysis between NSCLC-specific LLR6 and pan-cancer LLR6, as described above, specifically focusing on predicting the ICB response of participants with NSCLC (Extended Data Fig. 6a–c). Remarkably, the NSCLC-specific LLR6 consistently demonstrated higher AUCs across all datasets, with particularly notable improvements observed in the Shim et al. and Vanguri et al. cohorts (Extended Data Fig. 6a). Interestingly, the pan-cancer model demonstrated a robust capability to predict survival in participants with NSCLC, achieving impressively consistent HR values near 0.5 for both PFS and OS across various datasets (Extended Data Fig. 6b,c). Additionally, we also constructed and compared a simplified two-variable NSCLC-specific model (LLR2) using only TMB and PD-L1 TPS as input (Extended Data Fig. 6d–f). The LLR2 model exhibited slightly reduced AUC values on most datasets (Extended Data Fig. 6d); it also demonstrated compromised predictive capacity for patient survival, such as an inability to differentiate OS in the Vanguri et al. cohort (Extended Data Fig. 6f).

Furthermore, to evaluate the added value of PD-L1 TPS information, we applied the NSCLC-specific LLR6 model to other cancers, including gastric cancer, esophageal cancer and mesothelioma. Remarkably, even when tested across these distinct cancer types without further training or adaptation, the model still demonstrated superior predictive power for ICB response compared to using TMB or PD-L1 TPS alone (Fig. 7a), despite the limited sample size.

Additional analyses were performed to test our methodology's robustness. Firstly, the pan-cancer LLR6 model's accuracy persisted even when excluding NSCLC data (Extended Data Figs. 7a–d and 8a–d). Furthermore, retraining the model without using NSCLC data did not further improve its performance (Extended Data Fig. 9a). In addition, removing the cancer type term from the model, which adjusts for varying TMB loads and responses across cancers, slightly reduced predictive power but not significantly (Extended Data Fig. 9b). Lastly, systemic therapy history is typically not considered in ICB response prediction. Excluding this feature to predict response with a five-feature logistic LASSO model (LLR5) showed that both pan-cancer and NSCLC-specific LLR5 models performed slightly worse than their full counterparts, albeit not significantly (Extended Data Fig. 10a,b). Equations for computing pan-cancer and NSCLC-specific LORIS using LLR5 models are included below (Methods). These results underscore the robustness of our methodology.

In summary, we developed a logistic regression-based methodology, identifying key predictors such as TMB, systemic therapy history, albumin, NLR, age, cancer type (pan-cancer) and PD-L1 TPS (NSCLC) to estimate patient ICB response probabilities (Fig. 7b). LORIS is now publicly available at https://loris.ccr.cancer.gov to aid cancer immunotherapy researchers, clinicians and patients.

Discussion

The clinical utility of many machine learning models is markedly hindered by their black box nature, which makes them difficult to interpret[33–35]. To address this issue, we developed two interpretable models to predict patient ICB response: a pan-cancer model and an NSCLC-specific model. The clinical score derived here, LORIS, demonstrated a substantial improvement in predicting ICB response compared to current clinical biomarkers. Across multiple unseen datasets, the pan-cancer LORIS had a 15–68% increase in AUC over the TMB biomarker and the NSCLC-specific LORIS showed 4–17% and 5–23% increases in AUC over the PD-L1 and TMB biomarkers, respectively. Remarkably, despite using a limited set of features, LORIS matched or exceeded the performance of more complex computational methods, even under stringent testing conditions. For instance, when stratifying the Chowell et al. cohort by the year of initiating ICB therapy—a method considered stronger than random data splitting according to the TRIPOD guideline[36]—we observed that the pan-cancer LORIS demonstrated superior predictive power compared to the original approach[22]. This improvement was evident across both repeated cross-validation and unseen test data (Extended Data Fig. 2). Similarly, the NSCLC-specific LORIS achieved comparable performance to the original approach (AUC = 0.77 versus 0.80) on the Vanguri et al. cohort, despite the original approach using a notably larger feature set derived from radiology, histology and genomics and its performance being measured through cross-validation rather than using independent data[27]. These findings showcase the versatility of our approach, which is applicable to both pan-cancer studies and when tailoring robust models to specific cancer types.

From a translational perspective, our results demonstrated three key findings. Firstly, LORIS predicts not only patient ICB objective response but also short-term and long-term survival benefit following immunotherapy better than existing methods. More importantly, our model successfully identifies low-TMB or low-PD-L1 TPS patients who can still benefit from immunotherapy. Lastly, LORIS scores patients by their response probabilities to immunotherapy in a much more monotonic and consistent manner, leading to more accurate identification of likely responders and more effective exclusion of likely nonresponders. Taken together, LORIS could be a reliable tool for improving clinical decision-making practices in precision medicine to maximize patient benefit.

Notably, a patient's systemic therapy history, while not typically considered in ICB response prediction, had a significant role in both models. Theoretically, chemotherapy reduces immune system competency and could lead to reduced ICB response rates[37]. Indeed, it was shown that first-line chemotherapy can influence the tumor microenvironment and decrease the efficacy of subsequent immunotherapy[38]. It was also observed that resistance to anti-MAPK (mitogen-activated protein kinase) targeted therapy could promote an immune-evasive tumor microenvironment and cross-resistance to subsequent immunotherapy in melanoma cases[39]. More recently, it was found that removing systemic therapy history decreased the predictive power of ICB response[22]. However, a patient's systemic therapy history may also be influenced by multiple clinical factors guiding treatment decisions and, in time, may become less relevant as ICB drugs move into first-line therapy for more indications. We explored the impact of excluding this feature on prediction accuracy. Consequently, we found that its exclusion slightly compromised the model's predictive power; however, the effect was not significant (Extended Data Fig. 10a,b).

This study had a few limitations. Firstly, our study had a retrospective design; to further demonstrate the transformative value of LORIS in clinical settings, more prospective studies need to be conducted in the future. Secondly, although we curated a large cohort with comprehensive clinical, pathologic and genomic features measured in a single study, the sample size was still limited for most individual cancer types. As a result, we could build cancer-type-specific models only for NSCLC. Additionally, we did not have transcriptomic data for the participants, which is an important factor in assessing tumor microenvironment and predicting ICB response[21,40–43]. Similarly, we opted not to include detailed gene mutation or copy number alteration information in the current model because of ethical restrictions regarding the sharing of such data. The use of federated learning[44] is required for training and externally validating models that use this type of data. This, however, constitutes an independent research question and falls outside the scope of the current study. Lastly, the PD-L1 TPS data were mainly limited to participants with NSCLC and rarely measured in other cancer types. Despite this limitation, our preliminary analysis showed that the NSCLC-specific LORIS, which incorporates the PD-L1 TPS information, can also enhance the predictive power of ICB objective response in other cancer types (Fig. 7a). However, as the sample size was still very limited, further validation is needed to confirm the importance of PD-L1 expression in predicting ICB response in individual cancer types using more extensive cohorts.

In summary, this study analyzed a large and diverse cohort of participants with cancer treated with immunotherapy, including their clinical, pathologic and genomic data and ICB response information, which allowed us to develop a robust machine learning model to predict patients' objective response and survival following ICB therapy. LORIS integrates a few easily measurable patient features and produces monotonic scores, which have the potential to facilitate clinical decision-making and patient stratification (Fig. 7b). As our understanding of tumor immunology and the availability of comprehensive data in larger cohorts continue to improve, we expect to see the development of even more accurate models for personalized precision therapy, ultimately reducing cancer mortality.

Methods

Description of the ICB cohorts

The use of the participant data from the MSK1 and MSK2 cohorts was approved by the MSK institutional review board. All participants provided informed consent to a MSK IRB-approved protocol. All other cohorts were published previously. All participant features were collected before the start of ICB therapy. Covariate characteristics are summarized in Table 1 including sex, age, systemic therapy history, cancer type and treatment type.

Chowell et al. cohort. The Chowell et al. cohort comprised 1,479 participants diagnosed with 16 different types of solid tumors. The cohort data included measurements of 18 features. Sixteen of them were previously reported, including tumor information (MSI status, TMB, FCNA, HED and LOH in HLA-I), clinical information (sex, age, systemic therapy history before immunotherapy, BMI, cancer type, tumor stage and ICB drug class) and blood parameters (NLR and levels of albumin, platelets and hemoglobin). TMB was calculated as the total number of somatic nonsynonymous mutations in the tumor normalized to the exonic coverage of the respective MSK-IMPACT panels (in mutations per Mb). For more detailed information, please refer to ref. 22. Two additional features, that is, tumor PD-L1 TPS (available for a subset of participants) and the start year of receiving ICB therapy were extracted from the electronic health records for the purpose of this study and were not previously reported (Supplementary Table 6). PD-L1 TPS was determined using the Dako PD-L1 IHC 22C3 pharmDx kit (Agilent Technologies), which is approved by the FDA.

Shim et al. cohort. The Shim et al. cohort included 198 participants with advanced NSCLC, with 13 features measured. These features included tumor information (PD-L1 TPS, TMB and LOH in HLA-I), clinical information (sex, age, systemic therapy history before immunotherapy, smoking status, histology, Eastern Cooperative Oncology Group (ECOG) performance status and ICB drug class) and blood parameters (NLR and albumin levels). TMB was defined as the number of nonsynonymous alterations, identified from WES. PD-L1 TPS was assessed using the FDA-approved Dako PD-L1 IHC 22C3 pharmDx kit (Agilent Technologies) in the samples. For more detailed information, please refer to ref. 25. Among these features, blood NLR and albumin levels and participants' systemic therapy history were extracted from the electronic health records for the purpose of this study and were not previously reported (Supplementary Table 6).

MSK1 and MSK2 cohorts. The participants in the MSK1 and MSK2 cohorts were treated and their tumors were profiled with the MSK-IMPACT platform as part of standard clinical care. Participants selected for this study were those with solid tumors diagnosed from 2014 through 2019 who received at least one dose of ICB at MSK. We excluded participants with a history of more than one cancer, those without a complete blood count within 30 days before the first dose of ICB and those enrolled in blinded trials. We excluded participants who received ICB in a neoadjuvant or adjuvant setting and participants with an unevaluable response. The final set consisted of 557 participants with solid tumors from 17 different types. A total of 13 features were measured in the study, including tumor information (PD-L1 TPS (available for a subset of participants), TMB and FCNA), clinical information (sex, age, systemic therapy history before immunotherapy, cancer type, ICB drug class and the start year of receiving ICB therapy) and blood parameters (NLR and levels of albumin and platelets) (Supplementary Table 6). The measurement of clinical and genomic features was the same as for the Chowell et al. cohort.

Vanguri et al. cohort. The Vanguri et al. cohort included 247 participants with advanced NSCLC, with 15 features measured. One sample with unknown primary tumor site was excluded. These features included tumor information (PD-L1 TPS, TMB, FCNA and MSI status), clinical information (sex, age, systemic therapy history before immunotherapy, smoking status, tobacco use, histology, ECOG performance status, ICB drug class and the panels used for TMB determination) and blood parameters (NLR and albumin levels). TMB was calculated as the total number of somatic nonsynonymous mutations in the tumor normalized to the exonic coverage of the respective MSK-IMPACT panels (in mutations per Mb). PD-L1 immunohistochemistry was performed on 4-µm formalin-fixed paraffin-embedded tumor tissue sections using a standard PD-L1 antibody (E1L3N, dilution 1:100; Cell Signaling Technologies) validated in the clinical laboratory at the study institution. For more detailed information, please refer to ref. 27.

Kato et al. cohort. The Kato et al. cohort comprised 429 participants, with 35 participants from eight solid tumor types included in this study on the basis of three criteria: (1) participants received immunotherapy; (2) their cancer types were included in the Chowell et al. cohort; and (3) TMB was measured. Six features were assessed: tumor MSI status, TMB, sex, age, systemic therapy history before immunotherapy and cancer type. TMB was determined using panel next-generation sequencing performed by a CLIA (Clinical Laboratory Improvement Amendments)-certified laboratory. For more detailed information, please refer to ref. 26.

Ravi et al. cohort. The Ravi et al. cohort included 393 participants with NSCLC treated with anti-PD-L1 therapy, with 10 features measured; a total of 309 participants with TMB measured were included in this study. These features included tumor information (PD-L1 expression and TMB) and clinical information (sex, age, systemic therapy history before immunotherapy, tumor stage, smoking status, tobacco use, histology and ICB drug class). TMB was defined as the number of nonsynonymous alterations, identified from WES at the Genomics Platform of the Broad Institute of Harvard and MIT (Massachusetts Institute of Technology). For more detailed information, please refer to ref. 28.

Pradat et al. cohort. The Pradat et al. cohort comprised 1,031 participants with different types of cancer; a total of 57 participants from 13 solid tumor types were included in this study on the basis of three criteria: (1) participants received immunotherapy; (2) their cancer types were included in the Chowell et al. cohort; and (3) TMB was measured. Nine features were assessed, including tumor information (TMB and FCNA), clinical information (sex, age, systemic therapy history before immunotherapy, ICB drug class and cancer type) and blood parameters (NLR and albumin levels). TMB was determined using WES. For more detailed information, please refer to ref. 29.

Description of the MSK non-ICB cohort

This cohort comprised a subset of participants from a previously study[45]. In brief, our selection process focused on participants first diagnosed between 2015 and 2018, who presented with solid tumors that underwent NGS during MSK-IMPACT and subsequently received cancer therapy at MSK ($n = 14{,}577$). Subsequently, we excluded participants who had a history of more than one primary cancer ($n = 3{,}425$), those with cancer types comprising fewer than 100 cases ($n = 797$) and those with cancers of unknown primary origin ($n = 122$). Furthermore, participants who had ever received ICB treatment were also excluded from our analysis ($n = 2{,}022$)[45]. For more detailed information, please refer to ref. 45. Following these exclusion criteria, we selected participants who had complete data available for the six features used by the LLR6 model ($n = 4{,}872$). To ensure consistency with the number of participants in the combined Chowell test and MSK1 sets, which were used in Figs. 3 and 4, we randomly sampled participants within each cancer type. It is worth noting that, for certain cancer types, the total count of non-ICB-treated participants was lower than that in the combined Chowell test and MSK1 sets, where all available participants of that specific cancer type were included. The final dataset comprised 841 participants with solid tumors originating from 15 different types (Supplementary Table 6).

Dealing with missing data and extreme values

Blood NLR and albumin levels were not accessible for the Kato et al. and Ravi et al. cohorts. We input the average values of 3.8 and 6.2, respectively, from participants in the Chowell train set to represent these missing values for all participants in both cohorts. The NSCLC-specific LLR6 model included PD-L1 TPS as an input feature, which was not available for numerous participants. Consequently, only participants with available PD-L1 TPS data were included in the model's training and testing. A breakdown of the participant count with available PD-L1 TPS data across various cohorts is provided in Extended Data Fig. 1a. To mitigate the influence of extreme values in certain features, data truncation was implemented. Specifically, TMB values were truncated at 50 mutations per Mb, blood NLR was truncated at 25 and patient age was truncated at 85 years.

TMB and NLR harmonization

TMB for different ICB cohorts was determined using two different platforms, namely, WES (the Shim et al., Ravi et al. and Pradat et al. cohorts) and targeted tumor sequencing (other cohorts). We harmonized WES TMB values (TMB_{WES}, total mutation counts) to TMB values measured by the MSK-IMPACT targeted gene panel (TMB_{MSK}, mutations per Mb) based on the linear relationship derived previously[46]. Specifically, $TMB_{MSK} = 1.05 \times TMB_{WES}/S$, where S is the total length of the exons used for WES (in Mb).

Derived NLR (dNLR) was measured for the Vanguri et al. cohort. To harmonize it with the NLR in the merged MSK cohort, we used a strategy similar to that used for TMB standardization. In this case, an empirical relationship of NLR = 2 × dNLR was used on the basis of previous studies[47,48].

Patient outcomes following immunotherapy

Patient outcome following immunotherapy was assessed by measuring objective response, OS and PFS for all cohorts described above. Objective response was categorized on the basis of the RECIST version 1.1 criteria[30], except for the CNS tumors in the MSK2 cohort, where the Response Assessment in Neuro-Oncology (RANO) criteria were used instead[49]. The objective response was then dichotomized into responders (CR and PR) and nonresponders (SD and PD). PFS was defined as the time from the first infusion of ICB to disease progression or death from any cause. Participants without disease progression were censored at their last disease assessment. OS was defined as the time from the first ICB infusion to death from any cause and participants who were still alive at the time of review were censored at their last contact.

Developing multivariable models of ICB response

Pan-cancer study. Although randomly splitting a single dataset into model training and validation sets was used for developing RF16 (Chowell et al.) in ref. 22, it is believed to be a weak and inefficient form of validation, whereas splitting by time is a stronger approach[36]. Following the TRIPOD guideline, we used participants who underwent immunotherapy between 2015 and 2017 in the Chowell et al. cohort as our training set (Chowell train, $n = 964$) and used the other participants in the Chowell et al. cohort (who underwent immunotherapy in 2018; $n = 515$) as a test set. We investigated 20 machine learning classifiers to predict participant ICB response, using the FDA-approved TMB biomarker as a baseline model. Among these models, the decision tree and random forest classifiers directly took the raw feature values as input. For all other classifiers, all feature values were standardized by converting them to z-scores before inputting to the models. We built, tuned and evaluated all the multivariable machine learning models using the scikit-learn package (version 1.2.1) and pytorch-tabnet (version 4.1.0) in the Python programming language. We determined the optimal hyperparameter combination by using a random search approach with the RandomizedSearchCV function to maximize the AUC scores in a fivefold cross-validation of the training data. We determined the total number of different hyperparameter combinations for each model as the minimum of 10,000 and the total number of all possible combinations. The detailed combinations of hyperparameters and the identified optimal combination for each of the 20 machine learning classifiers are elaborated as follows:

(1) LR16: the 16-feature logistic regression classifier using all 16 features measured in the Chowell et al. cohort. All combinations: solver = 'saga', penalty = 'elasticnet', class_weight = 'balanced', l1_ratio from 0 to 1 (step size 0.1), max_iter from 100 to 1,000 (step size 100) and C from 10^{-3} to 10^{3} (logarithmic step size 1). Optimal combination: solver = 'saga', penalty = 'elasticnet', max_iter = 100, l1_ratio = 0.1, class_weight = 'balanced' and C = 0.01.

(2) LLR6: the six-feature logistic regression classifier using TMB, systemic therapy history, blood albumin level, blood NLR, age and cancer type. All combinations: same as above. Optimal combination: solver = 'saga', penalty = 'elasticnet', max_iter = 100, l1_ratio = 1, class_weight = 'balanced' and C = 0.1.

(3) LR5 (noTMB): the five-feature logistic regression classifier subtracting TMB from the previous model. All combinations: same as above. Optimal combination: solver = 'saga', penalty = 'elasticnet', max_iter = 100, l1_ratio = 0.4, class_weight = 'balanced' and C = 0.01.

(4) RF16 (Chowell et al.): the 16-feature random forest classifier with hyperparameters reported in ref. 22. Optimal combination: n_estimators = 1,000, max_depth = 8, min_samples_leaf = 20 and min_samples_split = 2.

(5) RandomForest: the 16-feature random forest classifier retrained using the protocol in this study. All combinations: n_estimators from 200 to 2,000 (step size 200), max_features from 0.1 to 0.9 (step size 0.1), max_depth from 3 to 10 (step size 1), min_samples_leaf from 2 to 30 (step size 2) and min_samples_split from 2 to 30 (step size 2). Optimal combination: n_estimators = 400, max_features = 0.1, max_depth = 9, min_samples_leaf = 2 and min_samples_split = 8.

(6) RF6: the six-feature random forest classifier trained using the same protocol as for the development of RF16 (Chowell et al.)[22]. All combinations: n_estimators from 100 to 1,000 (step size 100), max_depth from 2 to 20 (step size 2), min_samples_leaf from 2 to 20 (step size 2) and min_samples_split from 2 to 20 (step size 2). Optimal combination: n_estimators = 900, max_depth = 8, min_samples_leaf = 8 and min_samples_split = 20.

(7) DecisionTree: the decision tree classifier. All combinations: splitter = 'best' or 'random', max_features from 0.1 to 0.9 (step size 0.1), max_depth from 3 to 10 (step size 1), min_samples_leaf from 2 to 30 (step size 2), min_samples_split from 2 to 30 (step size 2) and ccp_alpha = 0, 0.5, 1, 10 or 100. Optimal combination: splitter = random, max_features = 0.7, max_depth = 7, min_samples_leaf = 8, min_samples_split = 2 and ccp_alpha = 0.

(8) GBoost: the GBoost classifier. All combinations: learning_rate = 0.01, 0.03, 0.05, 0.1, 0.3 or 0.5, n_estimators from 200 to 2,000 (step size 200), min_samples_split from 2 to 30 (step size 2), min_samples_leaf from 2 to 30 (step size 2), max_depth from 3 to 10 (step size 1) and max_features from 0.1 to 0.9 (step size 0.1). Optimal combination: learning_rate = 0.03, n_estimators = 200, min_samples_split = 12, min_samples_leaf = 4, max_depth = 6 and max_features = 0.1.

(9) AdaBoost: the AdaBoost classifier. All combinations: n_estimators from 200 to 2,000 (step size 200), learning_rate = 0.01, 0.05, 0.03, 0.1, 0.3, 0.5 or 1 and algorithm = 'SAMME' or 'SAMME.R'. Optimal combination: n_estimators = 1,000, learning_rate = 0.3 and algorithm = SAMME.

(10) HGBoost: the HGBoost classifier. All combinations: learning_rate = 0.01, 0.03, 0.05, 0.1, 0.3 or 0.5, max_iter from 200 to 2,000 (step size 200), min_samples_leaf from 2 to 30 (step size 2), max_depth from 3 to 10 (step size 1) and l2_regularization = 0 or from 10^{-4} to 10^{2} (logarithmic step size 1). Optimal combination: learning_rate = 0.03, max_iter = 600, min_samples_leaf = 16, max_depth = 10 and l2_regularization = 100.

(11) XGBoost: the XGBoost classifier. All combinations: min_child_weight = 1 or from 2 to 30 (step size 2), max_depth from 3 to 10 (step size 1), n_estimators = 100 or from 200 to 1,000 (step size 200), learning_rate = 0.01, 0.03, 0.05, 0.1, 0.3 or 0.5, colsample_bytree = 0.5, 0.8 or 1, colsample_bynode from 0.2 to 1 (step size 0.2) and colsample_bylevel from 0.2 to 1 (step size 0.2). Optimal combination: min_child_weight = 6, max_depth = 7, n_estimators = 400, learning_rate = 0.01, colsample_bytree = 0.8, colsample_bynode = 0.2 and colsample_bylevel = 1.

(12) LightGBM: the LightGBM classifier. All combinations: learning_rate = 0.001, 0.003, 0.005, 0.01, 0.03, 0.05, 0.1 or 0.3, max_depth from 3 to 10 (step size 1), n_estimators from 200 to 2,000 (step size 200), num_leaves from 10 to 100 (step size 10), colsample_bytree from 0.2 to 1 (step size 0.2) and min_data_in_leaf from 2 to 30 (step size 2). Optimal combination: learning_rate = 0.03, max_depth = 3, n_estimators = 200, num_leaves = 30, colsample_bytree = 0.8 and min_data_in_leaf = 30.

(13) SupportVectorMachine: the support vector machine classifier. All combinations: C from 10^{-5} to 10^{3} (logarithmic step size 0.5), gamma = 'scale' or 'auto' or from 10^{-4} to 10^{2} (logarithmic step size 0.5), kernel = 'rbf', max_iter = −1, 100 or 1,000, tol from 10^{-5} to 10^{-1} (logarithmic step size 0.5) and class_weight = none or 'balanced'. Optimal combination: C = $10^{2.5}$, gamma = $10^{-3.5}$, kernel = 'rbf', max_iter = 1,000, tol = 0.001 and class_weight = none.

(14) kNearestNeighbors: the *k*-nearest neighbors classifier. All combinations: n_neighbors from 2 to 60 (step size 2), weights = 'uniform' or 'distance', algorithm = 'auto', 'ball_tree', 'kd_tree' or 'brute', leaf_size from 2 to 30 (step size 2) and p from 1 to 10 (step size 1). Optimal combination: n_neighbors = 58, weights = 'distance', algorithm = 'brute', leaf_size = 20 and p = 1.

(15) TabNet: the TabNet deep neural network classifier. All combinations: max_epochs = 50, n_d = 24 or 32, n_a = n_d, n_steps = 3, 4 or 5, gamma = 1, 1.5 or 2, lambda_sparse = 0.0001, 0.001 or 0.01 and momentum = 0.3, 0.4 or 0.5. Optimal combination: max_epochs = 50, n_d = 32, n_a = 32, n_steps = 5, gamma = 1.5, lambda_sparse = 0.0001 and momentum = 0.5.

(16) MultilayerPerceptron (one layer): the multilayer perceptron classifier (one layer). All combinations: solver = 'sgd', 'lbfgs' or 'adam', learning_rate = 'constant', 'invscaling' or 'adaptive', max_iter = 100, 200, 500 or 1,000, hidden_layer_sizes from 2 to 40 (step size 1) in one hidden layer, activation = 'logistic', 'tanh', 'relu' or 'identity', alpha from 10^{-6} to 10^{-1} (logarithmic step size 1) and early_stopping = false or true. Optimal combination: solver = 'adam', learning_rate = 'adaptive', max_iter = 200, hidden_layer_sizes = 19, activation = 'tanh', alpha = 10^{-2} and early_stopping = false.

(17) MultilayerPerceptron (two layers): the multilayer perceptron classifier (two layers). All combinations: max_iter = 100, 200, 500 or 1,000, hidden_layer_sizes from 2 to 20 (step size 1) in two hidden layers, activation = 'logistic', 'tanh', 'relu' or 'identity', alpha from 10^{-6} to 10^{-1} (logarithmic step size 1) and early_stopping = false or true. Optimal combination: max_iter = 100, hidden_layer_sizes = (19, 19), activation = 'tanh', alpha = 10^{-5} and early_stopping = false.

(18) MultilayerPerceptron (three layers): the multilayer perceptron classifier (three layers). All combinations: hidden_layer_sizes from 2 to 20 (step size 1) in three hidden layers, activation = 'logistic', 'tanh', 'relu' or 'identity' and alpha from 10^{-6} to 10^{-1} (logarithmic step size 1). Optimal combination: hidden_layer_sizes = (6, 5, 6), activation = 'relu' and alpha = 10^{-1}.

(19) MultilayerPerceptron (four layers): the multilayer perceptron classifier (four layers). All combinations: same as above. Optimal combination: hidden_layer_sizes = (3, 17, 2, 4), activation = 'tanh' and alpha = 10^{-3}.

(20) GaussianProcess: the Gaussian process classifier. All combinations: kernel = none, 1.0 × kernels.RBF (1.0), 0.1 × kernels.RBF (0.1) or 10 × kernels.RBF (10), optimizer = 'fmin_l_bfgs_b' or none, max_iter_predict = 100, 500 or 1,000 and n_restarts_optimizer from 0 to 30 (step size 5). Optimal combination: kernel = 10 × RBF (length_scale = 10), optimizer = none, max_iter_predict = 100 and n_restarts_optimizer = 0.

After hyperparameter tuning, it was observed that the logistic regression model with six features had a LASSO penalty proportion of 100%, making it an LLR model. For ease of reference, we referred to this model as LLR6 throughout the paper. The hyperparameters that were optimal for LLR6 were used to train the NSCLC-specific model. The regression coefficients, which included the intercept, from the pan-cancer LLR6 model were obtained by averaging the corresponding values obtained from the 10,000 training iterations described earlier. No further adaptation was performed on the test data. The LLR6 score, calculated using this model, was referred to as LORIS. The formula for pan-cancer LORIS is explicitly given as follows:

$$\begin{aligned}\mathrm{LORIS} = \tfrac{1}{1+e^{-S}}S &= 0.0371 \times \min(\mathrm{TMB},\ 50) - 0.8775\\ &\times \mathrm{PSTH} + 0.5382 \times \mathrm{Albumin} - 0.033 \times \min(\mathrm{NLR}, 25)\\ &+ 0.0049 \times \min(\mathrm{Age}, 85) + \mathrm{CTCT} - 2.0886\end{aligned}$$

where PSTH is the participant's systemic therapy history, a binary variable that indicates whether the participant has received chemotherapy or targeted therapy before immunotherapy (1) or has not (0) and CTCT is the cancer type calibration term, which is equal to −0.3323 × bladder − 0.3323 × breast − 0.102 × colorectal − 0.0079 × endometrial + 0.55 × esophageal + 0.2306 × gastric + 0.0678 × head and neck − 0.1189 × hepatobiliary − 0.0086 × melanoma + 0.1255 × mesothelioma + 0.0008 × NSCLC − 0.052 × ovarian − 1.1169 × pancreatic + 0.5451 × renal + 0.0542 × sarcoma − 0.0033 × SCLC.

We also trained a five-feature logistic regression model (LLR5) without using a patient's systemic therapy history. The formula for LORIS calculated using this model is as follows:

$$\begin{aligned}\mathrm{LORIS\ (LLR5)} = \tfrac{1}{1+e^{-S}}S &= 0.0384 \times \min(\mathrm{TMB}, 50) + 0.5789\\ &\times \mathrm{Albumin} - 0.046 \times \min(\mathrm{NLR}, 25)\\ &+ 0.0087 \times \min(\mathrm{Age}, 85) + \mathrm{CTCT} - 3.0063\end{aligned}$$

where CTCT is equal to −0.3821 × bladder − 0.5696 × breast − 0.2294 × colorectal − 0.0646 × endometrial + 0.5489 × esophageal + 0.4488 × gastric + 0.0193 × head and neck − 0.1316 × hepatobiliary + 0.296 × melanoma + 0.076 × mesothelioma − 0.0005 × NSCLC − 0.1136 × ovarian − 1.3838 × pancreatic + 0.5527 × renal + 0.0277 × sarcoma − 0.06 × SCLC.

NSCLC study. The NSCLC-specific study aimed to replicate the pan-cancer study, with the difference that only participants with NSCLC were used for training and evaluating the models. We used the Chowell et al. cohort as training data, which included 324 participants with NSCLC with complete data (Extended Data Fig. 1a).

As previously mentioned, we used the optimal hyperparameters obtained from the pan-cancer study to train the NSCLC-specific LLR6 model. We followed a similar approach to the pan-cancer modeling approach, calculating coefficients and intercepts for NSCLC-specific LLR6 on the basis of the average values of 10,000 training iterations, with 80% of the training data randomly selected for each iteration. The formula for LORIS calculated using NSCLC-specific LLR6 is as follows:

$$\begin{aligned}&\text{NSCLC-specific LORIS} = \tfrac{1}{1+e^{-S}}S = 0.0353 \times \min(\mathrm{TMB}, 50)\\ &+ 0.0111 \times \mathrm{PDL1} - 0.375 \times \mathrm{PSTH} + 0.2924 \times \mathrm{Albumin}\\ &- 0.0103 \times \min(\mathrm{NLR}, 25) + 0 \ \times \min(\mathrm{Age}, 85) - 1.5593\end{aligned}$$

Again, we also trained a five-feature logistic regression model (LLR5) without using the participants' systemic therapy history. The formulation for LORIS calculated using this model is as follows:

$$\begin{aligned}&\text{NSCLC-specific LORIS (LLR5)} = \tfrac{1}{1+e^{-S}}S = 0.0362\\ &\times \min(\mathrm{TMB}, 50) + 0.013 \times \mathrm{PDL1} + 0.3311 \times \mathrm{Albumin}\\ &- 0.0101 \times \min(\mathrm{NLR}, 25)\ +\ 0.0001 \times \min(\mathrm{Age}, 85)\ -\ 1.9915\end{aligned}$$

Model performance evaluation

To evaluate the performance of the models, we used 2,000 repeats of fivefold cross-validation on the training data. During each cross-validation fold, 80% of the training data were used for model training and the remaining 20% were used for evaluation. We used multiple metrics, such as AUC, AUPRC, accuracy, F_1-score, Matthews correlation coefficient[50] and balanced accuracy[51], to quantify the predictive power of the models. To determine the optimal threshold for the predicted response probabilities computed by the model, we maximized Youden's index, defined as 'sensitivity + specificity − 1'. We finally ranked the performance of each model using the geometric mean of four metrics: AUC, AUPRC, accuracy and F_1-score. As overfitting is a common problem in supervised models, we calculated the difference between the performance scores on the training data and cross-validation data for each cross-validation fold to estimate the extent of overfitting for each model.

Statistical analyses

We conducted various statistical analyses using Python (version 3.9) and R (version 4.1) to investigate the relationships between different variables and ICB response. Spearman's rank test from the scipy package (version 1.10.1) was used to calculate correlation coefficients and raw P values among features measured on a continuous scale, which were then adjusted for Bonferroni correction. To compare the distributions of response probability generated by different models (for example, LLR6, RF6 and TMB) between responders and nonresponders, we used the Mann–Whitney U test. DeLong's test[52] was used for comparison of AUCs. The 95% CIs of AUCs were calculated using 1,000 bootstrapping replicates.

Survival analysis was performed using the R packages survminer (version 0.4.9) and survival version (3.3.1). We calculated HRs with 95% CIs and P values with univariable Cox proportional hazards regression using the coxph() function[53]. In pan-cancer analysis, to compare differences in half-year, 1-year, 2-year, 3-year, 4-year and 5-year survival probability between high-LORIS versus low-LORIS or high-TMB versus low-TMB groups, we used the paired Wilcoxon rank sum test. Multivariable analysis was performed with Cox proportional hazards regression in individual cancer types using the coxph() function, with adjustment for cancer type, age, drug class of ICB and year of ICB start. In NSCLC-specific analysis, multivariable analysis was performed with adjustment for sex, age and drug class of ICB.

To stratify participants on the basis of risk, two methods were used for each variable of interest: an absolute threshold (for all histologies) or a percentile threshold (within each histology). In the case of pan-cancer LORIS, the absolute and percentile thresholds were set at 0.5 and 50%, respectively. These thresholds were determined using the training data to maximize Youden's index for predicting ICB objective response. For NSCLC-specific LORIS, an absolute threshold of 0.44, also determined on the basis of the training data, was applied. Regarding TMB, absolute and percentile thresholds of 10 mutations per Mb and 80% (that is, the highest 20% within each histology) were used, following ref. 5. For PD-L1 TPS in NSCLC, an absolute threshold of 50% was used.

For variables such as LORIS, PD-L1 TPS and TMB, we calculated the average values and 95% CIs using 1,000 bootstrapping replicates of the data to determine their relationship with ICB objective response probability. The response rate of participants falling within the range of $x-0.05$ to $x+0.05$ was used as a surrogate to estimate the probability of patient objective response at each specific LORIS value x. Likewise, the intervals of $(x-5\%, x+5\%]$ and $(x-5, x+5]$ were used for PD-L1 TPS and TMB values, respectively. As all the features were normalized by z-score before being put into the model, the feature importance for the logistic regression model was directly shown as the absolute values of the corresponding coefficients in the model (Extended Data Fig. 1c,d).

For all statistical tests used, data distribution was not assumed to be normal. No statistical methods were used to predetermine sample sizes but our sample sizes are substantially larger than those reported in previous publications[22,25–29]. The cohorts were already randomized as they were participants in clinical trials. The investigators were blinded to the response annotations until they became available.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The original data of the Chowell et al. cohort are available in Supplementary Table 3 of ref. 22. The original data of the Shim et al. cohort are available in Supplementary Table 1 of ref. 25. The original data of the Vanguri et al. cohort are available at Synapse (https://www.synapse.org/#!Synapse:syn26642505) and cBioPortal (https://www.cbioportal.org/study/summary?id=lung_msk_mind_2020). The original data of the Kato et al. cohort are available in Supplementary Data 1 of ref. 26. The original data of the Ravi et al. cohort[28] are available on Zenodo (https://doi.org/10.5281/zenodo.7625517)[54]. The original data of the Pradat et al. cohort are available in the Supplementary Tables of ref. 29. Deidentified new data reported in this study for the MSK1 cohort, MSK2 cohort and MSK non-ICB cohort, as well as additional features of participants in the Chowell et al. and Shim et al. cohorts that have not been reported before are included in Supplementary Table 6 and are available online at Zenodo (https://doi.org/10.5281/zenodo.11186449)[55]. Source data are provided with this paper.

Code availability

All codes that are necessary to reproduce all the results in the paper are implemented in Python and R and are publicly available at GitHub (https://github.com/rootchang/LORIS)[56] and Zenodo (https://doi.org/10.5281/zenodo.11186449)[55].

References

1. Topalian, S. L., Taube, J. M., Anders, R. A. & Pardoll, D. M. Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy. *Nat. Rev. Cancer* **16**, 275–287 (2016).
2. Morad, G., Helmink, B. A., Sharma, P. & Wargo, J. A. Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. *Cell* **184**, 5309–5337 (2021).
3. Nishino, M., Ramaiya, N. H., Hatabu, H. & Hodi, F. S. Monitoring immune-checkpoint blockade: response evaluation and biomarker development. *Nat. Rev. Clin. Oncol.* **14**, 655–668 (2017).
4. Goodman, A. M. et al. Tumor mutational burden as an independent predictor of response to immunotherapy in diverse cancers. *Mol. Cancer Ther.* **16**, 2598–2608 (2017).
5. Samstein, R. M. et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nat. Genet.* **51**, 202–206 (2019).
6. McGrail, D. J. et al. High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. *Ann. Oncol.* **32**, 661–672 (2021).
7. Topalian, S. L. et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. *New Engl. J. Med.* **366**, 2443–2454 (2012).
8. Zhao, P. F., Li, L., Jiang, X. Y. & Li, Q. Mismatch repair deficiency/microsatellite instability-high as a predictor for anti-PD-1/PD-L1 immunotherapy efficacy. *J. Hematol. Oncol.* **12**, 54 (2019).
9. Mandal, R. et al. Genetic diversity of tumors with mismatch repair deficiency influences anti-PD-1 immunotherapy response. *Science* **364**, 485–491 (2019).
10. Le, D. T. et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409–413 (2017).
11. Chowell, D. et al. Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nat. Med.* **25**, 1715–1720 (2019).
12. Chowell, D. et al. Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* **359**, 582–587 (2018).

13. Davoli, T., Uno, H., Wooten, E. C. & Elledge, S. J. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, eaaf8399 (2017).
14. Chang, T. G. et al. Optimizing cancer immunotherapy response prediction by tumor aneuploidy score and fraction of copy number alterations. *npj Precis. Oncol.* **7**, 54 (2023).
15. Ren, F. P., Zhao, T., Liu, B. & Pan, L. Neutrophil–lymphocyte ratio (NLR) predicted prognosis for advanced non-small-cell lung cancer (NSCLC) patients who received immune checkpoint blockade (ICB). *Onco. Targets Ther.* **12**, 4235–4244 (2019).
16. Valero, C. et al. Pretreatment neutrophil-to-lymphocyte ratio and mutational burden as biomarkers of tumor response to immune checkpoint inhibitors. *Nat. Commun.* **12**, 729 (2021).
17. Yoo, S. K., Chowell, D., Valero, C., Morris, L. G. T. & Chan, T. A. Pre-treatment serum albumin and mutational burden as biomarkers of response to immune checkpoint blockade. *npj Precis. Oncol.* **6**, 23 (2022).
18. Wang, Z. M. et al. Paradoxical effects of obesity on T cell function during tumor progression and PD-1 checkpoint blockade. *Nat. Med.* **25**, 141–151 (2019).
19. Conforti, F. et al. Cancer immunotherapy efficacy and patients' sex: a systematic review and meta-analysis. *Lancet Oncol.* **19**, 737–746 (2018).
20. Kugel, C. H. et al. Age correlates with response to anti-PD1, reflecting age-related differences in intratumoral effector and regulatory T-cell populations. *Clin. Cancer Res.* **24**, 5347–5356 (2018).
21. Litchfield, K. et al. Meta-analysis of tumor- and T cell-intrinsic mechanisms of sensitization to checkpoint inhibition. *Cell* **184**, 596–614 (2021).
22. Chowell, D. et al. Improved prediction of immune checkpoint blockade efficacy across multiple cancer types. *Nat. Biotechnol.* **40**, 499–506 (2022).
23. Gromeier, M. et al. Very low mutation burden is a feature of inflamed recurrent glioblastomas responsive to cancer immunotherapy. *Nat. Commun.* **12**, 352 (2021).
24. Diggs, L. P. & Hsueh, E. C. Utility of PD-L1 immunohistochemistry assays for predicting PD-1/PD-L1 inhibitor response. *Biomark. Res.* **5**, 12 (2017).
25. Shim, J. H. et al. HLA-corrected tumor mutation burden and homologous recombination deficiency for the prediction of response to PD-(L)1 blockade in advanced non-small-cell lung cancer patients. *Ann. Oncol.* **31**, 902–911 (2020).
26. Kato, S. et al. Real-world data from a molecular tumor board demonstrates improved outcomes with a precision N-of-One strategy. *Nat. Commun.* **11**, 4965 (2020).
27. Vanguri, R. S. et al. Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer. *Nature Cancer* **3**, 1151–1164 (2022).
28. Ravi, A. et al. Genomic and transcriptomic analysis of checkpoint blockade response in advanced non-small cell lung cancer. *Nat. Genet.* **55**, 807–819 (2023).
29. Pradat, Y. et al. Integrative pan-cancer genomic and transcriptomic analyses of refractory metastatic cancer. *Cancer Discov.* **13**, 1116–1143 (2023).
30. Eisenhauer, E. A. et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
31. Cho, M. S. et al. Platelets increase the expression of PD-L1 in ovarian cancer. *Cancers* **14**, 2498 (2022).
32. Sechidis, K. et al. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics* **34**, 3365–3376 (2018).
33. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019).
34. Petch, J., Di, S. & Nelson, W. Opening the black box: the promise and limitations of explainable machine learning in cardiology. *Can. J. Cardiol.* **38**, 204–213 (2022).
35. Watson, D. S. et al. Clinical applications of machine learning algorithms: beyond the black box. *BMJ* **364**, l886 (2019).
36. Moons, K. G. M. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann. Intern. Med.* **162**, W1–W73 (2015).
37. Sambi, M., Bagheri, L. & Szewczuk, M. R. Current challenges in cancer immunotherapy: multimodal approaches to improve efficacy and patient response rates. *J. Oncol.* **2019**, 4508794 (2019).
38. He, Y. Y. et al. Genomic and transcriptional alterations in first-line chemotherapy exert a potentially unfavorable influence on subsequent immunotherapy in NSCLC. *Theranostics* **11**, 7092–7109 (2021).
39. Haas, L. et al. Acquired resistance to anti-MAPK targeted therapy confers an immune-evasive tumor microenvironment and cross-resistance to immunotherapy in melanoma. *Nat. Cancer* **2**, 693–708 (2021).
40. Auslander, N. et al. Robust prediction of response to immune checkpoint blockade therapy in metastatic melanoma. *Nat. Med.* **24**, 1545–1549 (2018).
41. Jiang, P. et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* **24**, 1550–1558 (2018).
42. Bareche, Y. et al. Leveraging big data of immune checkpoint blockade response identifies novel potential targets. *Ann. Oncol.* **33**, 1304–1317 (2022).
43. Liu, D. et al. Integrative molecular and clinical modeling of clinical outcomes to PD1 blockade in patients with metastatic melanoma. *Nat. Med.* **25**, 1916–1927 (2019).
44. Konečný, J. et al. Federated learning: strategies for improving communication efficiency. Preprint at https://doi.org/10.48550/arXiv.1610.05492 (2016).
45. Valero, C. et al. The association between tumor mutational burden and prognosis is dependent on treatment context. *Nat. Genet.* **53**, 11–15 (2021).
46. Merino, D. M. et al. Establishing guidelines to harmonize tumor mutational burden (TMB): in silico assessment of variation in TMB quantification across diagnostic platforms: phase I of the Friends of Cancer Research TMB Harmonization Project. *J. Immunother. Cancer* **8**, e000147 (2020).
47. Kim, C. G. et al. On-treatment derived neutrophil-to-lymphocyte ratio and survival with palbociclib and endocrine treatment: analysis of a multicenter retrospective cohort and the PALOMA-2/3 study with immune correlates. *Breast Cancer Res.* **25**, 4 (2023).
48. Proctor, M. J. et al. A derived neutrophil to lymphocyte ratio predicts survival in patients with cancer. *Br. J. Cancer* **107**, 695–699 (2012).
49. Wen, P. Y. et al. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. *J. Clin. Oncol.* **28**, 1963–1972 (2010).
50. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F_1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
51. Velez, D. R. et al. A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.* **31**, 306–315 (2007).
52. DeLong, E. R., DeLong, D. M. & Clarke-Pearson, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837–845 (1988).

53. Therneau, T. A package for survival analysis in S. *CRAN* https://CRAN.R-project.org/package=survival (2015).
54. Holton, M., Arniella, M., Ravi, A. & Getz, G. Genomic and transcriptomic analysis of checkpoint blockade response in advanced non-small cell lung cancer. *Zenodo* https://doi.org/10.5281/zenodo.7625517 (2023).
55. Chang, T. LORIS: a logistic regression-based immunotherapy-response score. *Zenodo* https://doi.org/10.5281/zenodo.11186449 (2024).
56. Chang, T. et al. LORIS: a logistic regression-based immunotherapy-response score. *GitHub* https://github.com/rootchang/LORIS (2024).

Acknowledgements

This research is supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research. This work used the computational resources of the NIH HPC Biowulf cluster (https://hpc.nih.gov).

Author contributions

T.-G.C., E.R. and L.G.T.M conceptualized and designed the study. T.-G.C., Y.C. and S.R.D. developed the machine learning models. T.-G.C., H.J.S., Y.C., S.R.D., S.-H.L., C.V., S.-K.Y., D.C., L.G.T.M. and E.R. acquired, analyzed or interpreted the data. All authors critically revised the manuscript for important intellectual content. E.R. and L.G.T.M. supervised the study.

Competing interests

E.R. is a cofounder of MedAware, Metabomed and Pangea Biomed (divested) and an unpaid member of Pangea Biomed's scientific advisory board. L.G.T.M. is listed as an inventor on intellectual property owned by MSK related to the use of TMB in cancer immunotherapy, unrelated to this work. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-024-00772-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43018-024-00772-7.

Correspondence and requests for materials should be addressed to Luc G. T. Morris or Eytan Ruppin.

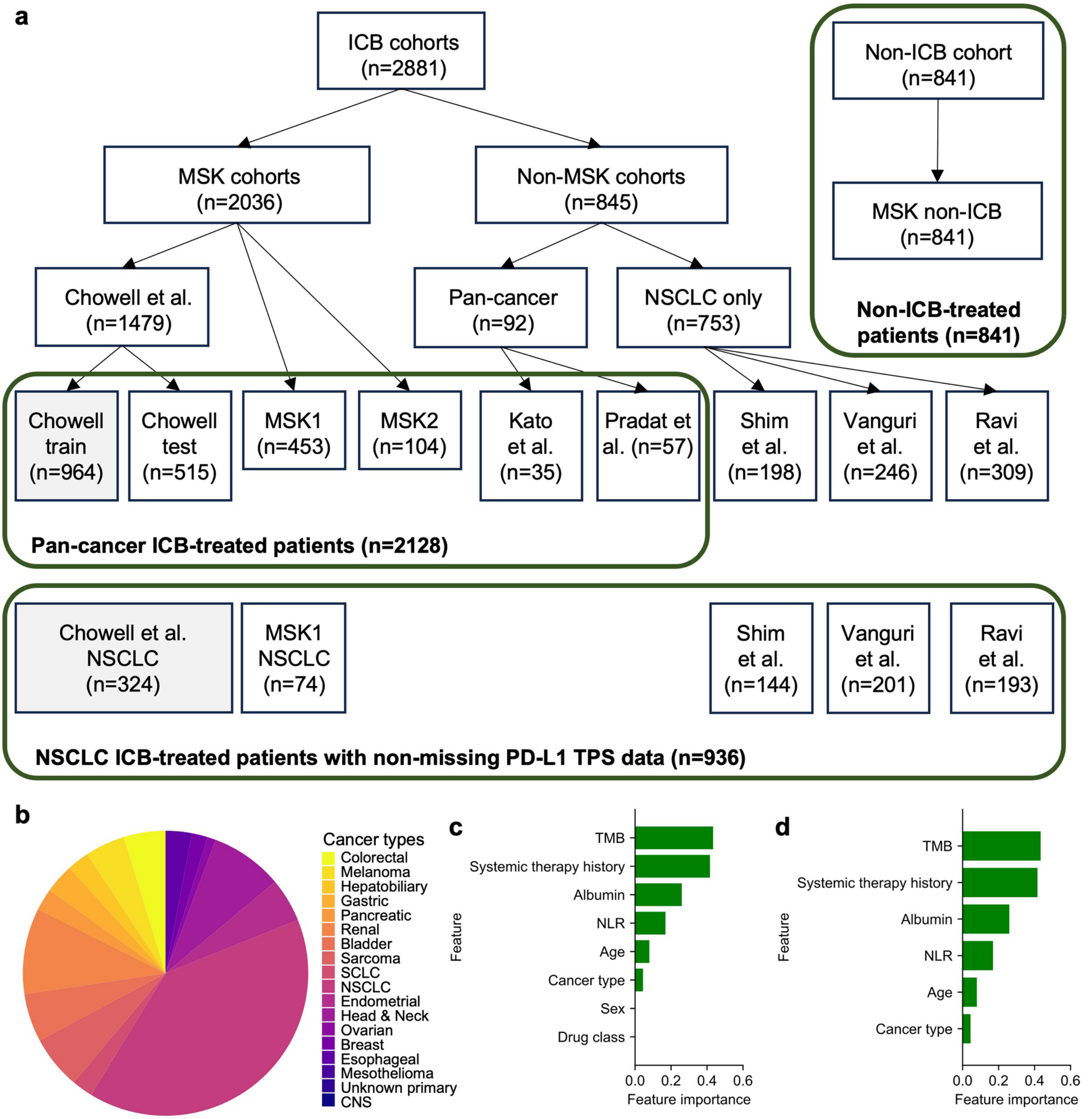
Peer review information *Nature Cancer* thanks Justin Gainor, Hajime Uno and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

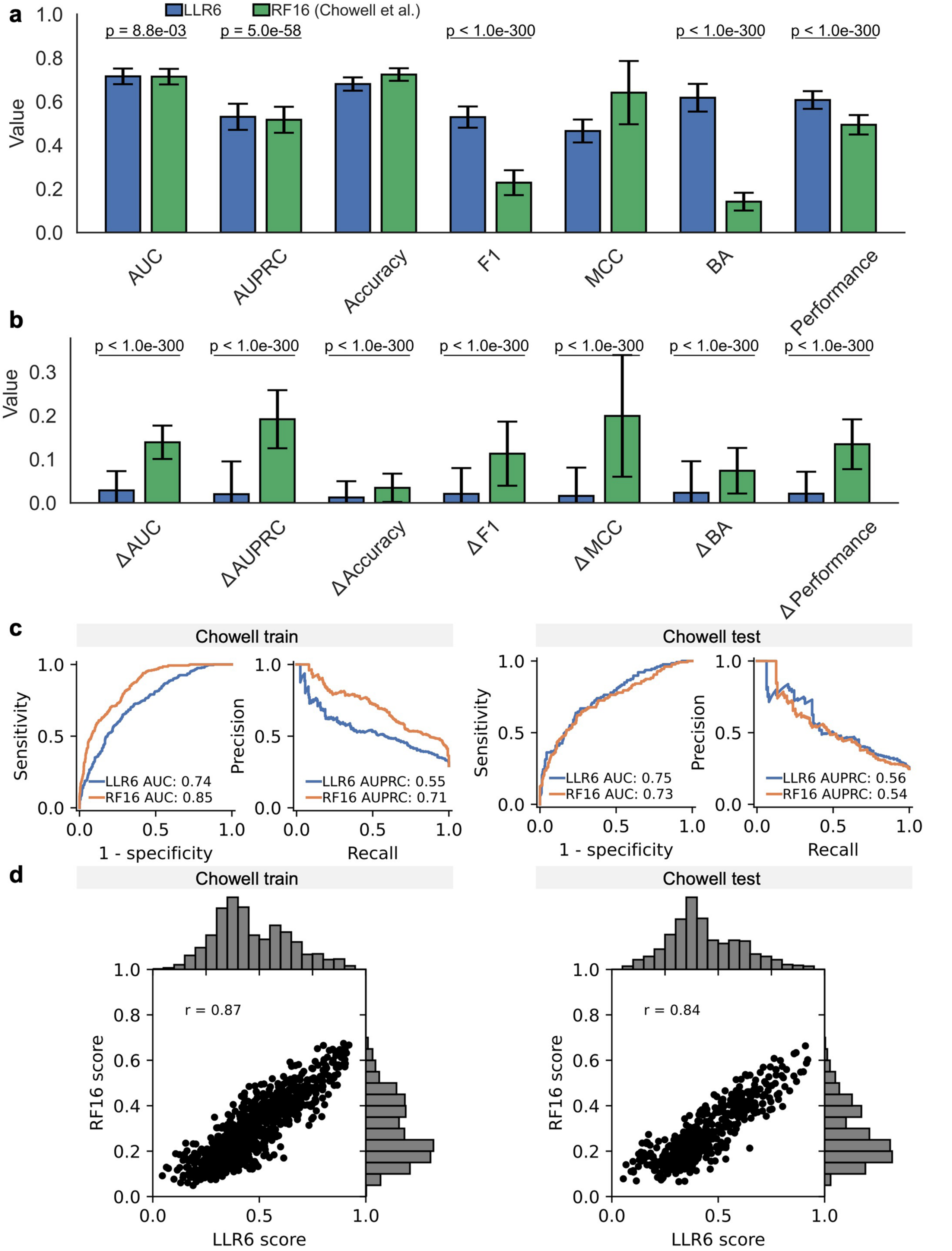
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2024

[1]Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, MD, USA. [2]Department of Surgery and Cancer Immunogenomics Research Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [3]Department of Health Sciences and Technology, Samsung Advanced Institute of Health Science and Technology, Sungkyunkwan University, Seoul, South Korea. [4]The Marc and Jennifer Lipschultz Precision Immunology Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]Department of Oncological Sciences, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]These authors contributed equally: Tian-Gen Chang, Yingying Cao, Hannah J. Sfreddo. ✉e-mail: morrisl@mskcc.org; eytan.ruppin@nih.gov



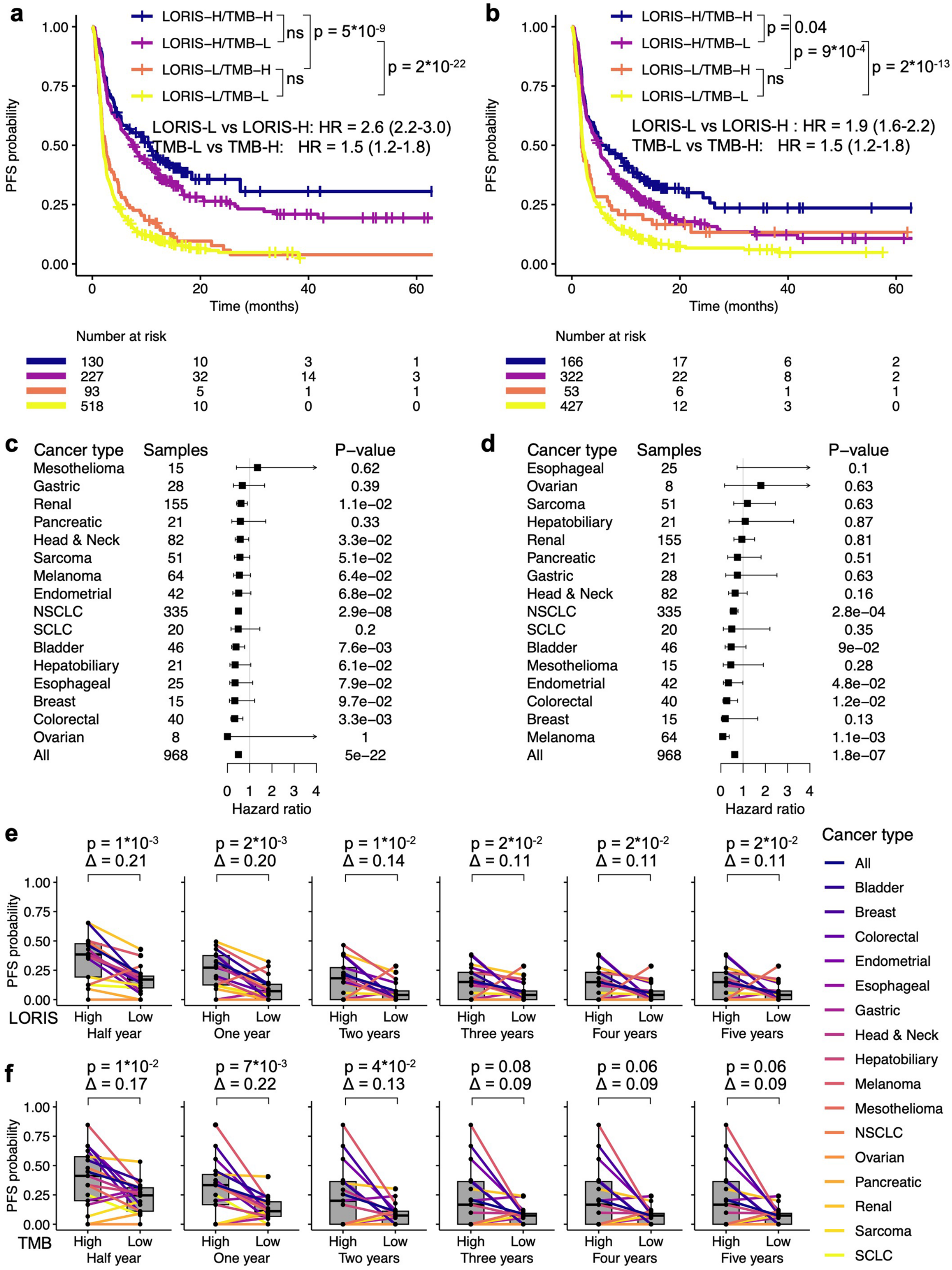
Extended Data Fig. 1 | An illustration of cohorts used in this study (a-b) and feature importance by the logistic LASSO regression model (c-d). a. The relationship between cohorts used in this study, the number of participants in each cohort, and the number of participants with complete data for the pan-cancer model and the NSCLC-specific model. The cohorts shaded in light grey represent the training cohorts for the pan-cancer and NSCLC-specific models, respectively. In the figure, 'n' represents the number of participants. **b**. The cancer composition of the non-ICB cohort. Note that three cancer types, mesothelioma, cancer of unknown primary, and central nervous system cancer are not present in this cohort. **c-d**. Feature importance of from the 8-feature logistic regression classifier using features commonly measured across most participants (**c**) and feature importance of the final 6-feature logistic regression classifier LLR6 (**d**). Feature importance is calculated as the absolute values of the corresponding coefficients in the logistic regression models. Importance for cancer type is calculated as the average importance of individual cancer types.



Extended Data Fig. 2 | See next page for caption.

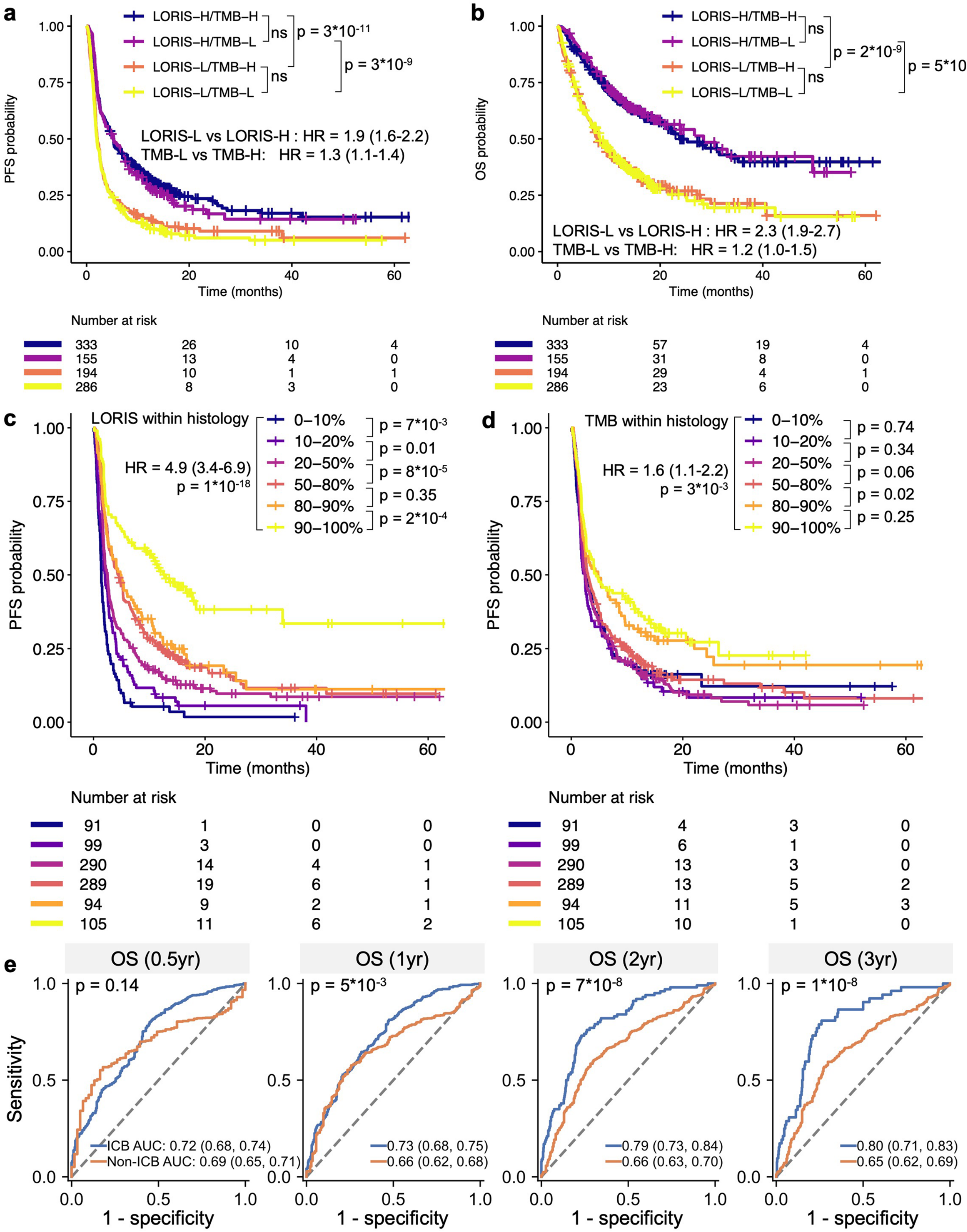
Extended Data Fig. 2 | Comparison between the pan-cancer LLR6 model and the RF16 (Chowell et al.) model. a. Comparison of the predictive power between the two models on 2,000-repeated 5-fold cross-validation sets using multiple metrics (n = 10,000 repetitions). Error bars, mean ± s.d. P values, two-tailed Mann-Whitney U test. Note that p values are only shown when values for LLR6 (blue bars) are significantly higher than RF16 (Chowell et al.) (green bars). **b**. Same as panel a, but the metrics represent the difference between those on the training sets and those on the corresponding cross-validation sets (n = 10,000 repetitions). Error bars, mean ± s.d. P values, two-tailed Mann-Whitney U test. **c**. Receiver operating characteristic curves and corresponding AUCs of LLR6 (blue curves) and RF16 (Chowell et al.) (orange curves) on the training (n = 964 participants) and unseen test (n = 515 participants) sets. Note that while the performance of RF16 (Chowell et al.) is better on the training set, the performance of the much simpler LLR6 model is better on the unseen test set. **d**. Correlation between the scores from LLR6 and RF16 (Chowell et al.) on both training and unseen test sets, respectively. Spearman correlation coefficients are shown.

https://doi.org/10.1038/s43018-024-00772-7



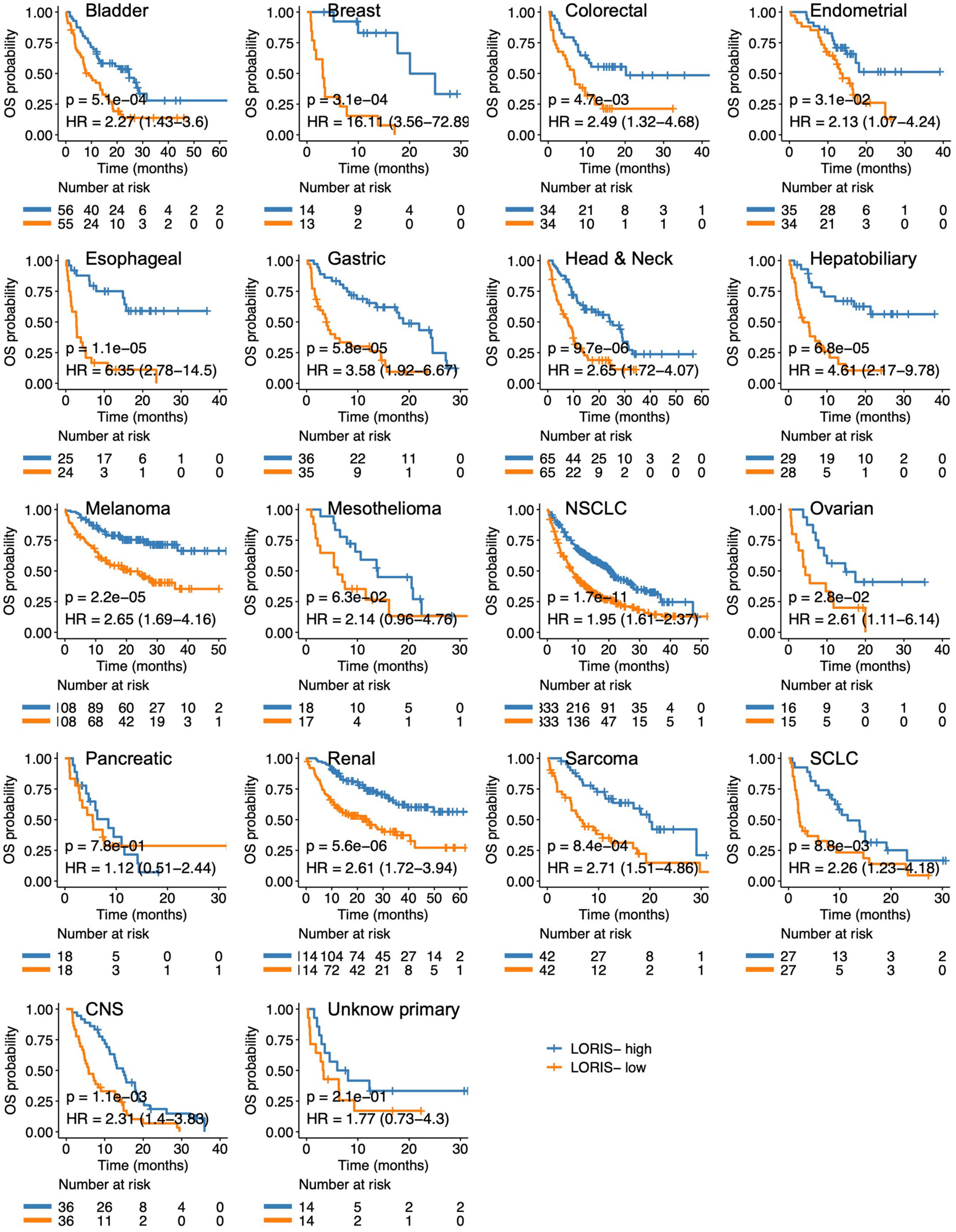
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | LORIS predicts PFS following immunotherapy for both pan-cancer and individual cancer types. a. Kaplan–Meier analysis of PFS. TMB is binned at 10 mutations per Mb and LORIS is binned at 0.5. HRs with 95% confidence intervals are shown. P values, univariable Cox proportional hazards regression. H, high; L, low. In the risk table, the numbers represent the number of participants. **b**. Same as panel a, but TMB is binned at the highest 20th percentile and LORIS is binned at the 50th percentile for each cancer type. HRs with 95% confidence intervals are shown. P values, univariable Cox proportional hazards regression. H, high; L, low. **c**, **d**. Forest plot of HRs of PFS within each cancer type using LORIS (binned at the 50th percentile; **c**) or TMB (binned at the highest 20th percentile; **d**). P values, multivariable Cox proportional hazards regression with adjustment for cancer type, age, ICB drug class, and year of ICB start. Squares positioned at midpoints symbolize point estimates of HRs, and the accompanying bars indicate 95% confidence intervals. **e**,**f**. Comparison of half-year, 1-year, 2-year, 3-year, 4-year, and 5-year PFS stratified by cancer type for high versus low LORIS (binned at the 50th percentile; **e**) and high versus low TMB (binned at the highest 20th percentile; **f**). Median survival probability differences (Δ) are displayed. P values, two-tailed paired Wilcoxon rank sum test. Box boundaries represent the first and third quartiles; the central line marks the median. Whiskers extend to the furthest non-outlier points within 1.5 times the interquartile range. Data are from combined *Chowell test* and *MSK1* sets (n = 968 participants).

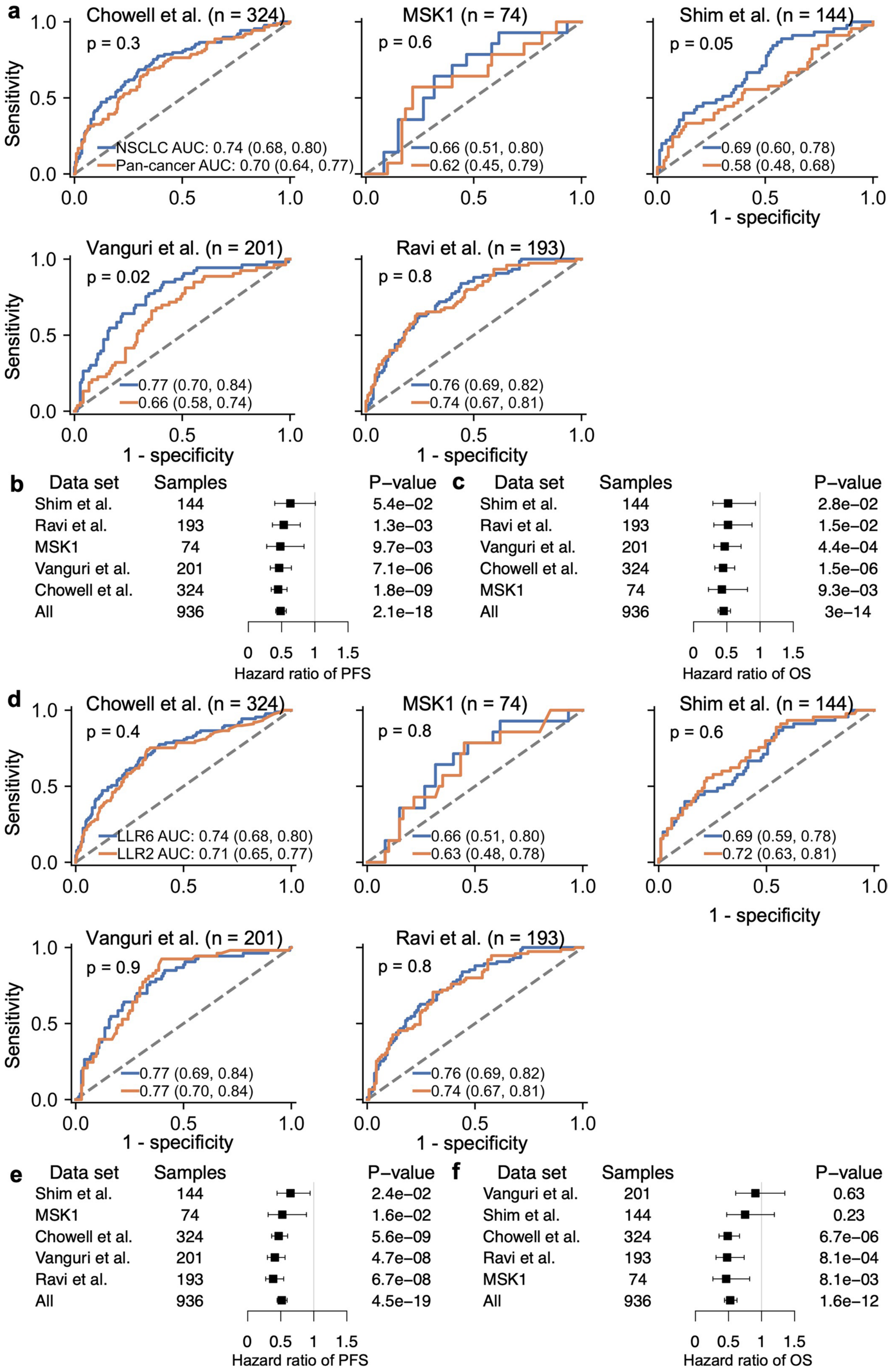


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | LORIS has better prediction power of immunotherapy than TMB (a-d) and has enhanced predictive power over prognosis (e). **a-b**. Kaplan–Meier analysis of PFS (a) and OS (b). Both TMB and LORIS are binned at the 50th percentile for each cancer type. HRs with 95% confidence intervals are shown. P values, univariable Cox proportional hazards regression. H, high; L, low. Data are from combined *Chowell test* and *MSK1* sets (n = 968 participants). **c-d**. Kaplan–Meier analysis of LORIS (**c**) or TMB (**d**) binned at the different percentiles in each cancer type. P values next to the legend indicate pairwise single-tail comparisons testing against the hypothesis that 'higher scored participants do not have better survival than lower scored participants' with univariable Cox proportional hazards regression. HRs with 95% confidence intervals are shown for the lowest-percentile (0–10%) and the highest-percentile groups (90–100%) with univariable Cox proportional hazards regression. Data are from combined *Chowell test* and *MSK1* sets (n = 968 participants). **e**. Receiver operating characteristic curves and corresponding AUCs with 95% confidence intervals of LORIS on 0.5-year OS, 1-year OS, 2-year OS, and 3-year OS of participants treated with ICB (blue curves) or non-ICB (orange curves) therapies. P values, two-tailed DeLong's test. ICB data are from combined *Chowell test* and *MSK1* sets (n = 968 participants). Non-ICB data are from the *MSK non-ICB* cohort (n = 841 participants). The dashed lines represent random performance, serving as a baseline with an AUC of 0.5. This indicates the performance expected from a classifier making random guesses.

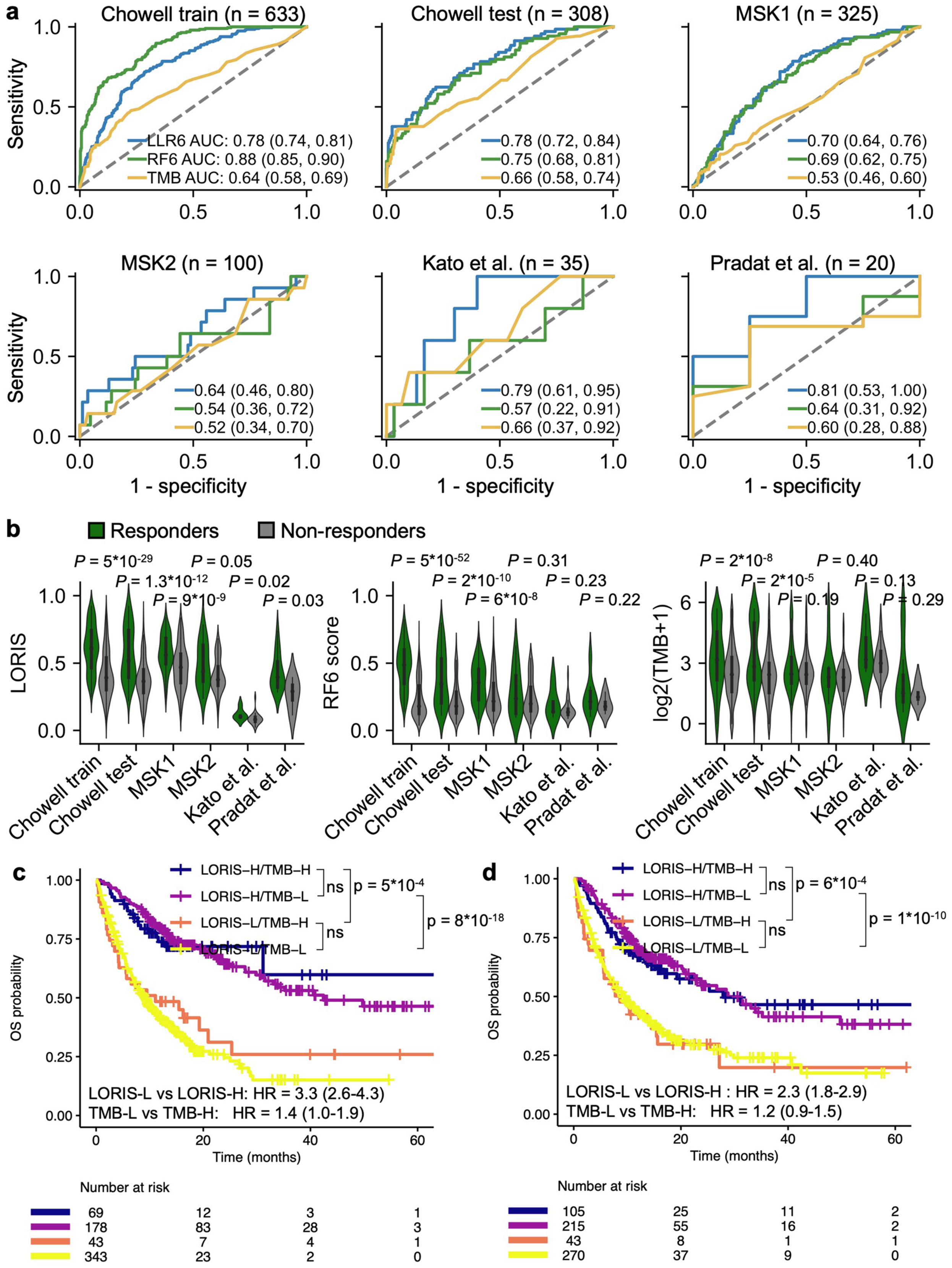


Extended Data Fig. 5 | Kaplan–Meier analysis of survival in individual cancer types. Patients are grouped into LORIS-high (orange curves) and LORIS-low (blue curves) risk groups. LORIS is binned at the 50th percentile for each cancer type. HRs with 95% confidence intervals are shown. P values, univariable Cox proportional hazards regression. In the risk tables, the numbers represent the number of participants. Data are from combined *Chowell* et al., *MSK1*, and *MSK2* sets (n = 2032 participants). Abbreviations: SCLC, small-cell lung cancer; CNS, central nervous system tumor; Unknown primary, cancer of unknown primary.



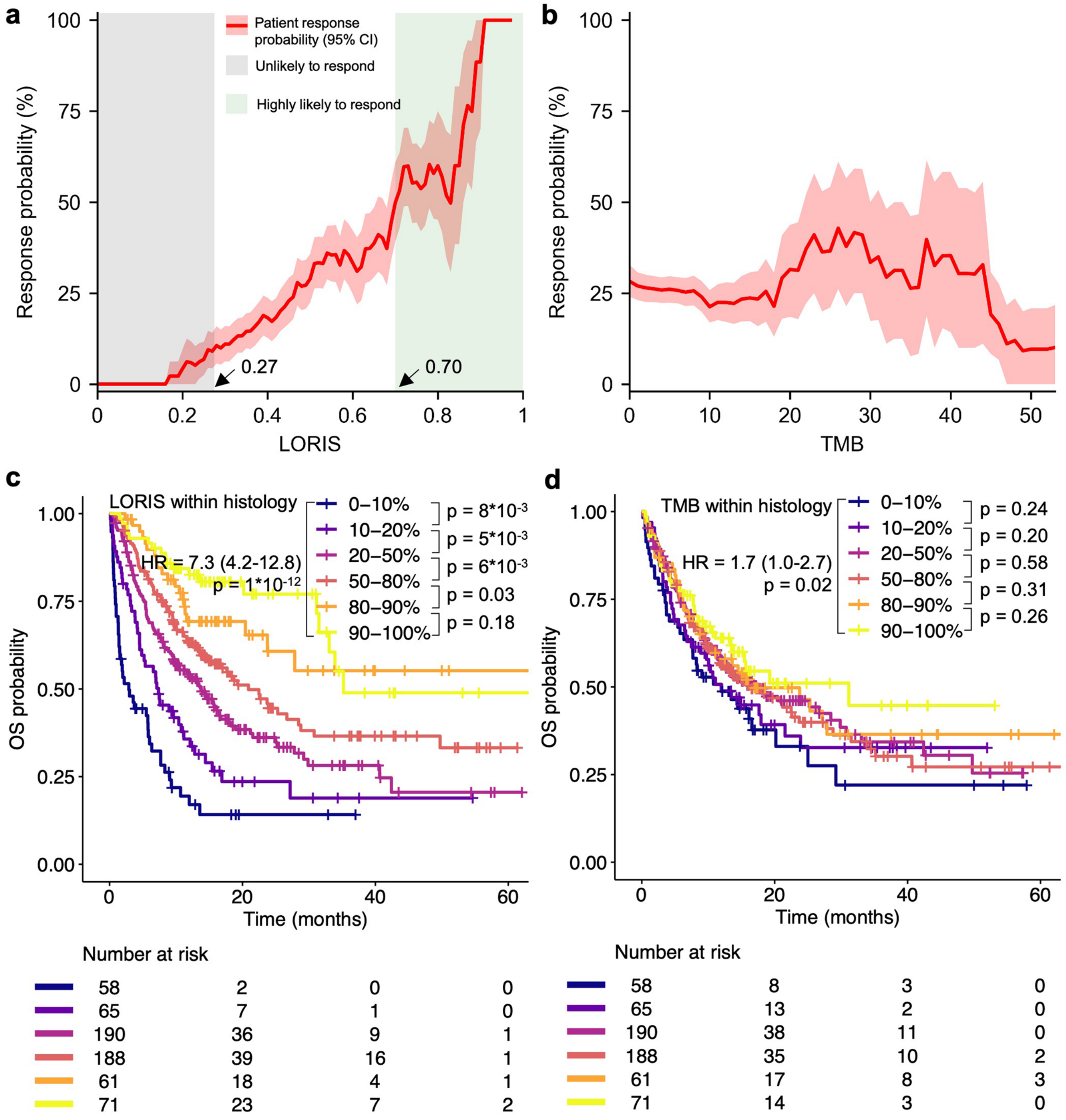
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Comparison of predictive performance between the NSCLC-specific LLR6, pan-cancer LLR6, and NSCLC-specific LLR2 models. **a**. Receiver operating characteristic curves and corresponding AUCs with 95% confidence intervals of the NSCLC-specific (blue curves) and pan-cancer (orange curves) LLR6 models. P values are from DeLong's test. In the figure, 'n' represents the number of participants. **b-c**. Forest plots of HRs of PFS (b) and OS (c) within each data set using pan-cancer LORIS (binned at 0.5, which maximizes the Youden's index on the training data) in a multivariable Cox model with adjustment for sex, age and ICB drug class. P values, multivariable Cox proportional hazards regression with adjustment for sex, age, and ICB drug class. Squares positioned at midpoints symbolize point estimates of HRs, and the accompanying bars indicate 95% confidence intervals. In the figure, the samples represent the number of participants. **d**. Receiver operating characteristic curves and corresponding AUCs with 95% confidence intervals of the LLR6 (blue curves) and LLR2 (orange curves) models. P values, two-tailed DeLong's test. The LLR2 model takes two variables, that is, patient TMB and PD-L1 TPS, as the input. In the figure, 'n' represents the number of participants. The dashed lines in **a** and **d** represent random performance, serving as a baseline with an AUC of 0.5. This indicates the performance expected from a classifier making random guesses. **e-f**. Forest plots of HRs of PFS (e) and OS (f) within each data set using LLR2 LORIS (binned at 0.46, which maximizes the Youden's index on the training data). P values, multivariable Cox proportional hazards regression with adjustment for cancer type and age. Squares positioned at midpoints symbolize point estimates of HRs, and the accompanying bars indicate 95% confidence intervals. In the figure, the samples represent the number of participants.

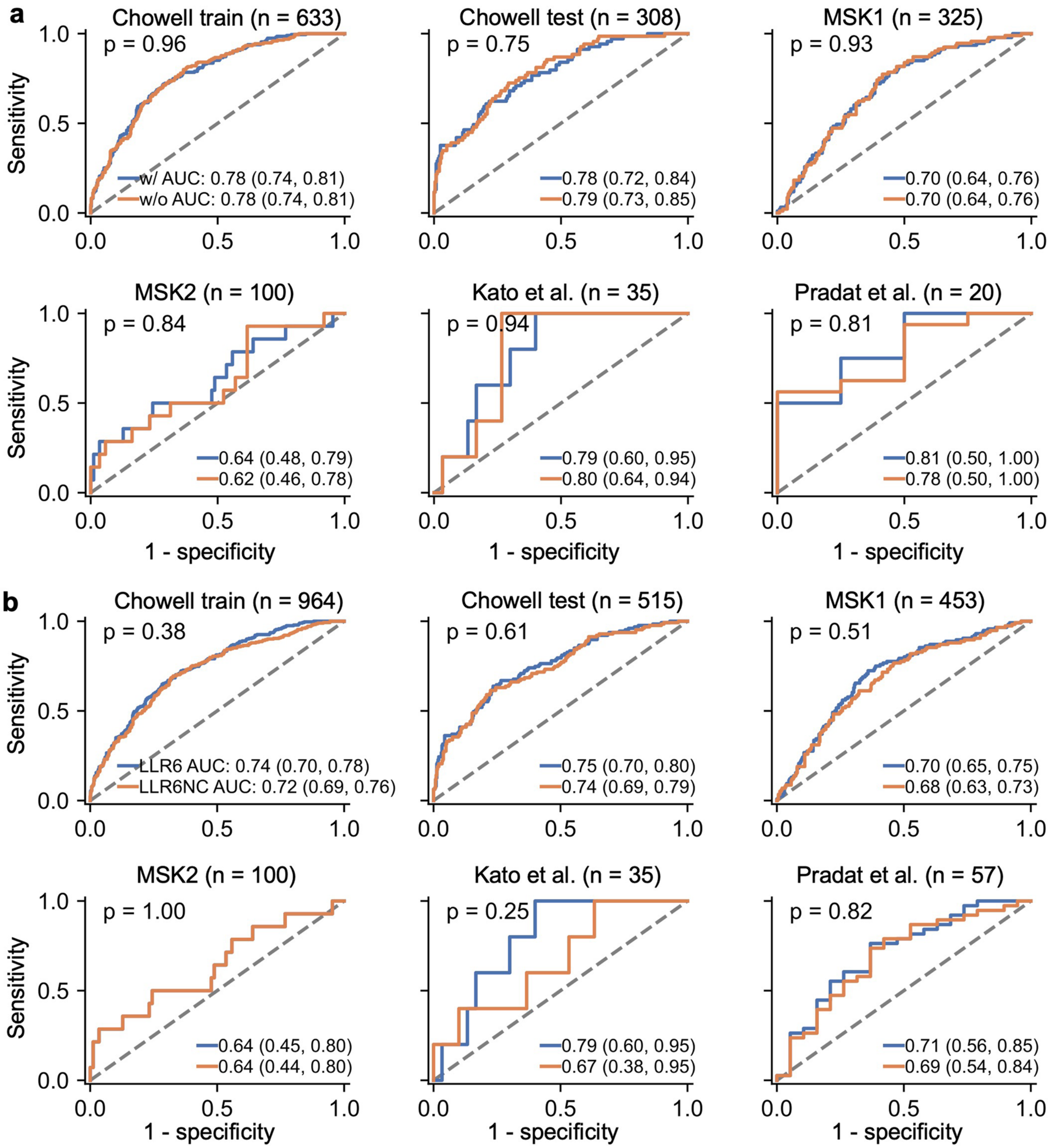


Extended Data Fig. 7 | See next page for caption.

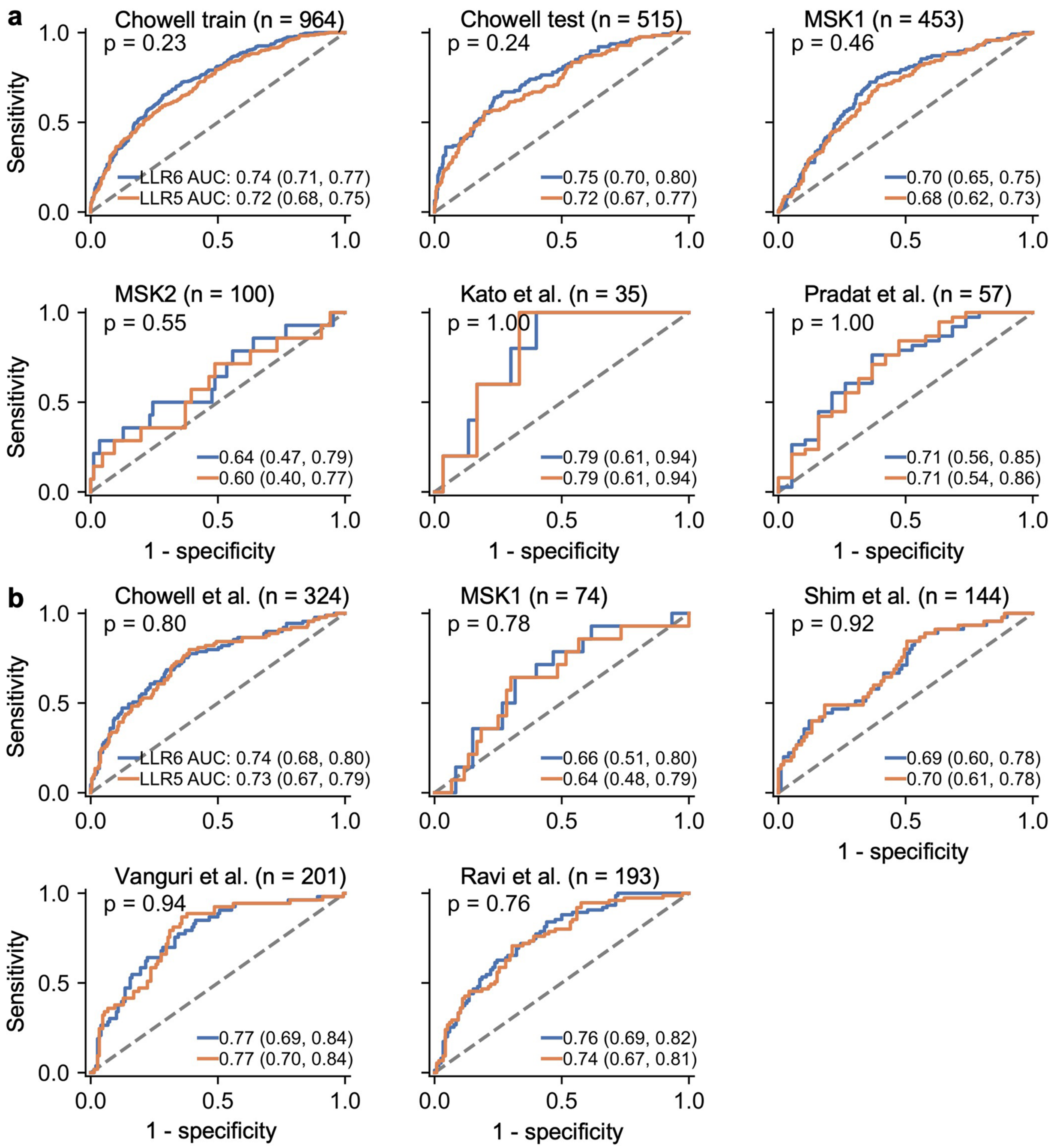
Extended Data Fig. 7 | Comparison of predictive performance of the pan-cancer LLR6 model, the RF6 model and TMB biomarker on non-NSCLC participants. a. Receiver operating characteristic curves and corresponding AUCs with 95% confidence intervals of LLR6 (blue curves), RF6 (green curves), and the TMB biomarker (yellow curves) on the training set and across multiple unseen test sets. In the figure, 'n' represents the number of participants. The dashed lines represent random performance, serving as a baseline with an AUC of 0.5. This indicates the performance expected from a classifier making random guesses. **b**. Distribution of LORIS, RF6 score, and TMB alone in responders and non-responders on the training set and across multiple unseen test sets. P values, two-tailed Mann–Whitney U test. Box boundaries represent the first and third quartiles; the central line marks the median. Whiskers extend to the furthest non-outlier points within 1.5 times the interquartile range. Outliers are shown as points beyond the whiskers. **c-d**. Kaplan–Meier analysis of OS. TMB is binned at 10 mutations per Mb and LORIS is binned at 0.5 for panel c; TMB is binned at the highest 20th percentile and LORIS is binned at the 50th percentile for each cancer type for panel d. HRs with 95% confidence intervals are shown. P values, univariable Cox proportional hazards regression. H, high; L, low. In the risk tables, the numbers represent the number of participants. Data are from combined *Chowell test* and *MSK1* sets, with all NSCLC patients excluded from the analysis (n = 633 participants).



Extended Data Fig. 8 | Monotonic relationship between pan-cancer LORIS and patient objective response probability & survival following immunotherapy among non-NSCLC participants. a, b. Relationship between LORIS (**a**) or TMB (**b**) and ICB objective response probability. The average participant response probabilities with 95% confidence intervals are shown using 1,000-replicate bootstrapping. The grey region represents participants with an unlikely response to immunotherapy (with a response probability below 10%), while the green regions represent participants with a likely response (with a response probability exceeding 50%). The arrows indicate the LORIS and TMB threshold values. **c, d**. Kaplan–Meier analysis of OS. LORIS (**c**) and TMB (**d**) are binned at the different percentiles in each cancer type. P values next to the legend indicate pairwise single-tail comparisons testing against the hypothesis that 'higher scored participants do not have better survival than lower scored participants' with univariable Cox proportional hazards regression. HRs with 95% confidence intervals are shown for the lowest-percentile (0–10%) and the highest-percentile groups (90–100%) with univariable Cox proportional hazards regression. In the risk tables, the numbers represent the number of participants. Data are from combined *Chowell test* and *MSK1* sets, with all NSCLC participants excluded from the analysis (n = 633 participants).



Extended Data Fig. 9 | LORIS performance is maintained after removing NSCLC participants (a) or removing cancer type information (b). **a**. Comparison of predictive performance among non-NSCLC participants between the original pan-cancer LLR6 model and a new LLR6 model trained without including NSCLC participants. Receiver operating characteristic curves and corresponding AUCs with 95% confidence intervals of the original pan-cancer LLR6 model (w/; blue curves) and a new LLR6 model trained without including NSCLC participants (w/o; orange curves). Number of participants in different cohorts is displayed in the figure. In the figure, ‘n’ represents the number of participants. P values, two-tailed DeLong’s test. Note that all NSCLC participants are excluded from the analysis. **b**. Comparison of predictive performance between the pan-cancer LLR6 model with and without the utilization of the cancer type calibration term. Receiver operating characteristic curves and corresponding AUCs with 95% confidence intervals of the original pan-cancer LLR6 model (LLR6; blue curves) and; orange curves). Number of participants in different cohorts is displayed in the figure. In the figure, ‘n’ represents the number of participants. P values, two-tailed DeLong’s test. The dashed lines represent random performance, serving as a baseline with an AUC of 0.5. This indicates the performance expected from a classifier making random guesses.



Extended Data Fig. 10 | Comparison of predictive performance between the LLR6 models and the LLR5 models that exclude a patient's systemic therapy history. a. Receiver operating characteristic curves and corresponding AUCs with 95% confidence intervals of the pan-cancer LLR6 (blue curves) and LLR5 (orange curves) models. Number of participants in different cohorts is displayed in the figure. In the figure, 'n' represents the number of participants. P values, two-tailed DeLong's test. **b.** Receiver operating characteristic curves and corresponding AUCs with 95% confidence intervals of the NSCLC-specific LLR6 (blue curves) and LLR5 (orange curves) models. Number of participants in different cohorts is displayed in the figure. In the figure, 'n' represents the number of participants. P values, two-tailed DeLong's test.

nature portfolio

Corresponding author(s): Eytan Ruppin, Luc Morris

Last updated by author(s): Jan 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	-All statistics were performed in R v4.1 and Python 3.9. -Kaplan-Meier survival analysis was performed using the R packages survminer v0.4.9 and survival v3.3.1. -Spearman's rank test was used to calculate correlation coefficients and raw p values between features measured on a continuous scale using the Python package scipy v.1.10.1, which were then adjusted for Bonferroni correction using the Python package statsmodels v0.13.5. -All machine learning models were built using the Python packages sklearn v1.2.1 and pytorch-tabnet v4.1.0.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The original data of Chowell et al. cohort are available in Supplementary Table 3 of the original publication (https://www.nature.com/articles/s41587-021-01070-8#MOESM1). The original data of Shim et al. cohort are available in Supplementary Table 1 of the original publication (https://www.annalsofoncology.org/article/S0923-7534(20)39295-4/fulltext#supplementaryMaterial). The original data of Vanguri et al. cohort are available at synapse: https://www.synapse.org/#!Synapse:syn26642505 and cbioportal: https://www.cbioportal.org/study/summary?id=lung_msk_mind_2020. The original data of Kato et al. cohort are available in Supplementary Data 1 of the original publication (https://www.nature.com/articles/s41467-020-18613-3#Sec16). The original data of Ravi et al. cohort are available at the Zenodo repository: https://doi.org/10.5281/zenodo.7625517. The original data of Pradat et al. cohort are available in Supplementary Tables of the original publication (https://aacrjournals.org/cancerdiscovery/article/13/5/1116/726168/Integrative-Pan-Cancer-Genomic-and-Transcriptomic).
De-identified new data reported in this study for the MSK1 & MSK2 cohorts, the MSK non-ICB cohort, and additional features of patients in the Chowell et al. and Shim et al. cohorts that have not been reported before are included in Supplementary Table 6 and are available online at Zenodo (https://doi.org/10.5281/zenodo.10679834).
All codes that are necessary to reproduce all the results in the paper are implemented in Python and R and are publicly available at GitHub (https://github.com/rootchang/LORIS) and Zenodo (https://doi.org/10.5281/zenodo.10679834).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	These findings are applicable to individuals of all sexes and genders, which were self-reported. The biological sex of participants is detailed in Table 1, with no significant gender association observed in the study.
Reporting on race, ethnicity, or other socially relevant groupings	This was not considered.
Population characteristics	Analyses were performed on 2881 ICB-treated patients from multiple cohorts and 841 non-ICB-treated patients from Memorial Sloan Kettering Cancer Center. Primary and metastatic patients across a broad range of histologies and age ranges were selected based on whether they received the indicated treatments. Covariate characteristics are summarized in Table 1 including sex, age, systemic therapy history, cancer type and treatment type.
Recruitment	This study was done retrospectively and no patients were directly recruited.
Ethics oversight	The use of the patient data from the MSK1 and MSK2 cohorts was approved by the Memorial Sloan Kettering Cancer Center institutional review board. All patients provided informed consent to a Memorial Sloan Kettering IRB-approved protocol. All other cohorts were published previously.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For cohorts other than MSK1 and MSK2, no sample size calculation were performed because they were available public data with predefined sample counts in each publication. The use of the patient data from MSK1 and MSK2 cohorts was approved by the MSKCC institutional review board. Patients selected for this study were those with solid tumors diagnosed from 2014 through 2019 who received at least 1 dose of ICB at MSKCC. The sample size was based on all available patients.
Data exclusions	For the Vanguri et al. cohort, 1 sample with unknown primary tumor site was excluded. For the Kato et al. and Pradat et al. cohorts, samples were selected based on three criteria: (1) patients received immunotherapy, (2) their cancer types are included in the Chowell et al. cohort, and (3) TMB was measured.

For the Ravi et al. cohort, samples without TMB measured were excluded.
For the MSK1 and MSK2 cohorts, we excluded patients with a history of more than 1 cancer, those without a complete blood count within 30 days prior to the first dose of ICB, those enrolled in blinded trials, and cancer types with fewer than 25 cases. We excluded patients who received ICB in a neoadjuvant or adjuvant setting, and patients with unevaluable response.

Replication

This study is retrospective, consisting of computational analyses using existing clinical datasets. Therefore, replication is not applicable since the counts were predetermined and cannot be altered by the authors. Instead, we collected multiple clinical datasets and used a part of data from one dataset as training set and used the other unseen data as test sets. The test sets were used to evaluate the model performance. Biological replicates are defined for each dataset in the figure legends.

Randomization

This is a retrospective study. The cohorts of patients were already randomized as they were participants in clinical trials.

Blinding

The investigators were blinded to the response annotations until they became available.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
□	☒	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern
☒	□	Plants

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Antibodies

Antibodies used

IHC was performed on 4-µm FFPE tumor tissue sections using a standard PD-L1 antibody (E1L3N; dilution 1:100, Cell Signaling Technologies, Danvers, MA).

Validation

The PD-L1 antibody was validated at the study institution according to manufacturer instructions.