Integration of clinical, pathological, radiological, and transcriptomic data improves the prediction of first-line immunotherapy outcome in metastatic non-small cell lung cancer

Authors and affiliations:

Nicolas Captier[1,2], Marvin Lerousseau[2,3], Fanny Orlhac[1], Narinée Hovhannisyan-Baghdasarian[1], Marie Luporsi[1,4], Erwin Woff[1,5], Sarah Lagha[6], Paulette Salamoun Feghali[6], Christine Lonjou[2], Clément Beaulaton[7], Hélène Salmon[8], Thomas Walter[2,3], Irène Buvat[1], Nicolas Girard[6], Emmanuel Barillot[2]

1: Laboratoire d'Imagerie Translationnelle en Oncologie, Institut Curie, Inserm U1288, PSL Research University, Orsay, France

2: Bioinformatics and computational systems biology of cancer, Institut Curie, Inserm U900, PSL Research University, Paris, France

3: CBIO-center for Computational Biology, MINES ParisTech, PSL Research University, Paris, France

4: Department of medical imaging, Institut Curie, Paris, France

5: Department of Nuclear Medicine/PET-scan, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium

6: Institut du Thorax Curie-Montsouris, Institut Curie, Paris, France

7: Department of pathology, Institut Curie, Paris, France

8: Immunity and cancer, Institut Curie, Inserm U932, PSL Research University, Paris, France

Corresponding authors:

Nicolas Captier - nicolas.captier[at]polytechnique.org
Emmanuel Barillot – emmanuel.barillot[at]curie.fr

Conflict of interest disclosure statement:

Nicolas Girard is employed by AstraZeneca (I) and has a consulting or advisory role for the following companies: Abbvie, AMGEN, AstraZeneca, BeiGene, Bristol-Myers Squibb, Daiichi Sankyo/Astra Zeneca, Gilead Sciences, Ipsen, Janssen, LEO Pharma, Lilly, MSD, Novartis, Pfizer, Roche, Sanofi, Takeda.

The other authors declare no potential conflict of interest.

Abstract:

The survival of patients with metastatic non-small cell lung cancer (NSCLC) has been increasing with immunotherapy, yet efficient biomarkers are still needed to optimize patient care. In this study, we explored the benefits of multimodal approaches to predict immunotherapy outcome using multiple machine learning algorithms and integration strategies. We leveraged a novel multimodal cohort of 317 metastatic NSCLC patients treated with first-line immunotherapy, collecting at baseline positron emission tomography images, digitized pathological slides, bulk transcriptomic profiles, and clinical information. Most integration strategies investigated yielded multimodal models surpassing both the best unimodal models and established univariate biomarkers, such as PD-L1 expression. Additionally, several multimodal combinations demonstrated improved patient risk stratification compared to models built with routine clinical features only. Our study thus provided new evidence of the superiority of multimodal over unimodal approaches, advocating for the collection of large multimodal NSCLC cohorts to develop and validate robust and powerful immunotherapy biomarkers.

Anti PD-1/PD-L1 immunotherapy with or without chemotherapy is the current standard first-line therapy for metastatic non-small cell lung cancer (NSCLC) without actionable oncogene alterations and without contraindications to PD-1/PD-L1 inhibitors (1). Several clinical trials have indeed demonstrated significantly improved Overall Survival (OS) and Progression-Free Survival (PFS) with immunotherapy in comparison to chemotherapy alone (2-6). Nevertheless, half of the patients do not present a radiological response to immunotherapy and the duration of response remains highly variable from one patient to another (ranging from 1.1 to 18 months for patients treated with first-line immunotherapy + chemotherapy) (3). Ultimately, the number of patients with long-term survival is limited. There is thus a critical need for biomarkers that can predict treatment response accurately. These biomarkers will pave the way to better personalize the treatment strategy – immunotherapy as single agent for patients with predicted prolonged survival, combination with chemotherapy or other agents for patients with predicted poor response and survival -, to customize the follow-up and assess adequately treatment sequences.

Machine learning approaches have recently shown their potential to leverage data collected before treatment initiation, including clinical (7, 8), radiological (9, 10), anatomopathological (11, 12), or transcriptomic information (13, 14), and develop robust prognostic and predictive models that could outperform approved univariate biomarkers such as PD-L1 expression (15). Promising results have fostered the exploration of multimodal approaches to combine all the diverse aspects of the disease that these different modalities probe. Yet, evidence of the superiority of multimodal over unimodal biomarkers (16) remains limited, possibly due to challenges in gathering comprehensive multimodal cohorts. Therefore, there is a pressing need for new studies involving large and homogeneous NSCLC multimodal cohorts to fully explore the benefits of multimodality and design strategies to address the challenges associated with integrating multimodal data.

In this study, we conducted a thorough comparison of the predictive capabilities of unimodal and multimodal approaches for predicting the outcome of NSCLC patients undergoing immunotherapy. We investigated a unique multimodal cohort of 317 metastatic NSCLC patients treated with first-line pembrolizumab, collecting four baseline data modalities: clinical data, PET/CT scans, digitized pathological slides, and bulk RNA-seq data. We developed methods to map each of these modalities to a set of interpretable features. Interpretability is a prerequisite to establish confidence in the predictive models and has the potential to reveal the most influential features for the prediction, thus enabling the discovery of a new generation of multimodal biomarkers. Subsequently, we trained models for each modality individually to predict OS and PFS, serving as a baseline for further analyses. We then built multimodal models, based on a variety of algorithms and settings, and compared them to their unimodal counterparts. Finally, we conducted feature importance analysis to identify important factors for immunotherapy outcome available in each modality.

Results

Clinical characteristics of patients with metastatic NSCLC.

We identified 317 NSCLC patients, treated at Institut Curie, who met the inclusion criteria: patients with histologically proven advanced NSCLC who received anti-PD-(L)1 immunotherapy, specifically pembrolizumab, as their first-line treatment. Immunotherapy was administered either as a standalone treatment for patients with a PD-L1 expression greater than 50%, or in combination with chemotherapy, regardless of the PD-L1 expression, as per clinical practice guidelines (1). The patients who received pembrolizumab as monotherapy were treated between October 2017 and January 2023 while those who received pembrolizumab combined with chemotherapy were treated between July 2019 and January 2023. The clinical characteristics of the multimodal cohort are detailed in Table 1.

Median OS and PFS were respectively 723 days (95% CI [446-987]) and 301 days (95% CI [145-598]) for the patients treated with immunotherapy alone, and 763 days (95% CI [576-NR]) and 290 days (95% CI [241-372]) for the patients treated with a combination of immunotherapy and chemotherapy (Figure 1.A). Interestingly, no significant difference was observed between the two treatment groups for OS

(log-rank p-value = 0.44, Figure 1.B), even for the patients with PD-L1 expression greater than 50% only (Figure s1). We observed that for PFS, the immunotherapy + chemotherapy group had fewer early progressors, although this was compensated by an increase in late progressors compared to the immunotherapy-only group (Figure 1.A).

Standard univariate biomarkers show limited predictive power.
PD-L1 expression was able to stratify patients, with significant differences in PFS and OS in patients with negative expression (< 1%) from those with positive expression (≥ 1%) (Figure 1.B, Figure s2). However, it yielded mild performance as a univariate biomarker for patient survival (C-index OS = 0.54, bootstrap 95% CI [0.51-0.57], permutation p-value = 0.014, n = 298). Besides, no significant performance was observed when using PD-L1 expression as a continuous score with the Tumor Proportion Score (TPS), where negative expressions were replaced by 0% and the score was calculated as 100 - TPS (C-index OS = 0.53, bootstrap 95% CI [0.48-0.58], permutation p-value = 0.104, n = 295). Other standard clinical biomarkers, such as the Tumor Mutational Burden (TMB) or Tumor Infiltrating Lymphocytes (TILs) did not exhibit significant association with patient outcome (Figure 1.C, Figure s3).

Collection of multiple baseline modalities to predict immunotherapy outcome.
Clinical information from routine care, 18F-FDG PET/CT scans, digitized pathological slides from the initial diagnosis, and bulk RNA-seq profiles from solid biopsies were collected at baseline. For each data modality, we first selected and computed several hand-crafted features to serve as input for both unimodal and multimodal predictive models, including 30 clinical features, 30 radiomic features, 134 pathomic features, and 34 transcriptomic features. We then leveraged this multimodal dataset to conduct an extensive comparison of the performance of unimodal and multimodal approaches using a 10-fold cross-validation scheme applied to the entire cohort and repeated 100 times (Figure s4).

237 out of the 317 patients had at least one missing modality (Figure 1.D). To ensure a fair comparison of all possible modality combinations, we therefore restricted the evaluation of prediction performance to the 80 patients with a complete multimodal profile (i.e., collecting the predictions of these 80 patients only, from the test sets of the cross-validation scheme applied to the whole cohort; Figure s4). Log-rank tests indicated no significant differences between the survival distributions of patients with missing modalities and those with available modalities (Figure s5).

Comparison of unimodal performances across multiple prediction tasks.
We first evaluated the predictive value of each modality individually. This evaluation involved predicting risk scores for time-to-event outcomes (OS and PFS) and classifying patients into two groups: those who would die within one year of treatment and those who would not (1-year death), or those who would experience disease progression before 6 months and those who would not (6-months progression). We focused on two standard Machine Learning approaches that are well-suited for datasets with modest numbers of samples (17): linear methods (logistic regression and Cox regression with elastic net penalties) and tree ensemble methods (gradient-boosted tree (18) and Random Survival Forest (19) algorithms). The four modalities exhibited varying degrees of predictive power for patient outcome, with the RNA modality standing out for the prediction of 1-year death (AUC = 0.75±0.04 (±1std); Table 2). PFS and 6-months progression predictions were more challenging than OS and 1-year death predictions. Except for pathological data, all modalities yielded greater performance (using either linear or tree ensemble algorithms) in predicting OS and 1-year death compared to PFS and 6-months progression. Across all modalities and models, the highest scores achieved were a C-index of 0.59 (±0.02) for RNA modality in predicting PFS and an AUC of 0.61 (±0.03) for clinical, pathomic, or RNA modality in predicting 6-months progression.

Feature importance analyses highlight relevant clinical and transcriptomic features.
We first investigated feature importance for the prediction of OS and 1-year death, providing insights into the information learned by each unimodal model (see Methods). Notably, it revealed that clinical

models consistently learned that patients with a low level of serum albumin, a negative PD-L1 status (i.e., TPS < 1%), or abundant circulating neutrophils were more likely to have a poor prognosis (Figure 2.A). This analysis also highlighted the transcriptomic features that the RNA models used to predict OS and 1-year death (Figure 2.B). RNA models consistently associated an abundance of dendritic cells (DCs), as scored by the MCP-counter method (20), or a high expression of NTRK1 gene with a good prognosis while they associated high expression of NRAS and KRAS genes with a poor prognosis. Interestingly, among the 13 consensus transcriptomic features identified with feature importance analysis (see Methods), only 3 exhibited significantly different values between biopsy sites in the 84 patients for whom this information was available (Figure s6), suggesting that the other 10 may be used independently of the biopsy location. Radiomic models primarily focused on information related to the Total Metabolic Tumor Volume (TMTV) as well as to the total metabolic volume of extra-thoracic metastases (Figure s7). Lastly, the interpretation of pathomic models unveiled features that encoded the proportion of inflammatory cells within the biopsy sections as well as their spatial organization (Figure s8).

We then turned to feature importance analysis of unimodal models for the prediction of PFS and 6-months progression. Notably, it confirmed that a high expression of the NTRK1 gene and an abundance of dendritic cells were associated with a favorable prognosis, since they were also ranked among the top ten most important features for multivariate predictions and showed significant univariate association with both PFS and 6-months progression (Figure s9.A). Similarly, it highlighted the favorable influence of positive PD-L1 expression or a high level of serum albumin on the prognosis of multivariate models, and the negative influence of a high neutrophils-to-lymphocytes ratio (Figure s9.B). Finally, this analysis showed that, similarly to OS and 1-year death prediction, radiomic models were predominantly driven by the TMTV (Figure s10.A).

Late fusion of unimodal predictors improves the prediction of immunotherapy outcome.
We then developed multimodal predictors with the hypothesis that multimodal data would provide richer and more comprehensive information. We first applied late fusion as a baseline strategy for integrating all the modalities into multimodal predictors of OS, 1-year death, PFS, and 6-months progression (Figure s1). Late fusion consists in averaging the predictions of each individual unimodal predictor. We tested every possible combination of two to four modalities for each predictive task using both linear and tree ensemble algorithms. The late fusion of tree ensemble models improved the prediction of patient outcome across both classification and survival tasks (Figure 3, Figure s11). Specifically, for 1-year death, the combination of predictions from clinical, RNA, and radiomic models demonstrated the highest performance (AUC = 0.81±0.03) while, for OS, the combination of predictions from clinical and RNA models performed best. For both OS and 1-year death, paired-permutation tests confirmed the significantly higher AUC and C-index for the combined model compared to clinical only, radiomics only, and pathomics only models (Figure s12). For PFS, the combination of predictions from clinical, RNA, pathomic, and radiomic models yielded the best performance while, for 6-months progression, it was the combination of predictions of clinical, RNA, and pathomic models. However, the performance of these two combinations was not significantly different from those of unimodal models with paired-permutation tests. The late fusion of linear models performed better than tree ensemble models for the prediction of 6-months progression only, with a combination of clinical, pathomic, and RNA predictions yielding an AUC of 0.67 (±0.03) (Figure s13). This can be explained by the greater performance of unimodal models with linear approaches for 6-months progression prediction, underscoring that the performance of late fusion combinations strongly depends on the performance of their unimodal components.

Benchmark of integration strategies reveals a consistent benefit of multimodal approaches.
We compared the late fusion approach with early fusion (Figure s1). The baseline early fusion approach consists in concatenating the features from the different modalities and using these concatenated vectors as input to a single predictor. For binary classification tasks, we also re-implemented and tested an

attention-based fusion approach known as DyAM (16), which was recently applied to NSCLC multimodal data. Early fusion and DyAM models were trained both without and with prior univariate feature selection to balance the dimension of the different modalities. The comparison of these different integration strategies for predicting OS, 1-year death, PFS, and 6-months progression did not identify a single best strategy (Figure 4). The late fusion of tree ensemble models yielded the best performance for the prediction of OS and 1-year death, while the early fusion of tree ensemble models and the DyAM model, both with prior univariate feature selection, outperformed the other strategies for PFS and 6-months progression prediction, respectively. This comparison demonstrated the potential of multimodal approaches to enhance unimodal predictions, as for each prediction task the majority of integration strategies resulted in multimodal combinations that outperformed the best unimodal models. Furthermore, this comparison highlighted modalities that were consistently involved in the best multimodal combinations across the different integration strategies, particularly for 1-year death and 6-months progression prediction. For 1-year death, the integration strategies that outperformed the best unimodal model with their optimal combination combined clinical (5/7 strategies), RNA (7/7 strategies), and radiomic (3/7 strategies) modalities. For 6-months progression, they combined clinical (7/8 strategies), pathomic (7/8 strategies), and RNA (8/8 strategies) modalities. The combination of clinical and RNA modalities also performed best for OS prediction, while for PFS prediction, it was the combination of clinical, pathomic, and RNA. Lastly, the superiority of multimodal approaches was confirmed when comparing the average performance of the different integration strategies across all possible combinations of one, two, three, and four modalities (Figure 5, Figure s14). Indeed, the average performance at a fixed number of modalities increased with the number of integrated modalities for every strategy and every prediction task (except for the early fusion with a linear model and no prior feature selection). Student's t-tests showed that multimodal combinations (involving two, three, or four modalities) consistently achieved significantly better average performance than unimodal models, except for PFS prediction. The performances of all the multimodal models are detailed in the supplementary materials (Figure s13, s15-22).

Multimodal predictions demonstrate improved patient stratification for OS.

Kaplan-Meier analysis showed that the predictions of multimodal models, integrating clinical data with other modalities when available, effectively stratified patients' OS. After adjusting the log-rank p-values, 93% of all the combinations across the different prediction tasks (i.e., 328/352) exhibited significant differences between the survival distributions of their low-risk and high-risk groups (Figure 6.A, Figure s23). Notably, 74% of the combinations (i.e., 260/352) yielded a lower p-value than the binary PD-L1 status (log-rank p-value = 0.0025, n=265). For each model, low-risk and high-risk group membership was defined by optimizing the cutoff on the training set of each cross-validation fold to maximize the log-rank test statistic and then applying that cutoff to the corresponding test set. In the case of classification tasks, a 0.5 cutoff on the predicted probabilities was also considered (see Methods). The clinical modality alone effectively separated patients into two risk groups, with the predictions of a linear model trained to predict 1-year death yielding the best p-value (log-rank p-value = 1.26e-06, Figure 6.B). For all prediction tasks, a range of one to seven multimodal models, out of the 56 possible models (i.e., integration strategy + multimodal combination), demonstrated superior risk stratification (as measured by the log-rank test statistic) compared to the best clinical model (Figure 6.A, Figure s23). Specifically, a combination of clinical, pathomic, and RNA modalities, trained to predict 1-year death with a tree ensemble algorithm, yielded the best p-value (log-rank p-value = 3.51e-09, Figure 6.B). We further compared the multimodal score resulting from this combination (i.e., the test predictions averaged across the different cross-validation schemes) with the clinical score derived from the linear model described above (Figure 6.B). To do so, we divided the cohort into quartiles based on these two scores and performed Kaplan-Meier analysis for OS within the first year of therapy (Figure 6.C). Both scores yielded a lowest quartile group (low risk) with a 14% death rate (9/66 patients) within the first year of therapy. However, the multimodal score identified a highest quartile group (high risk) with a 52% death rate (35/67 patients, 20 treated with immunotherapy + chemotherapy and 15 treated with

immunotherapy only), whereas the clinical score yielded a highest quartile group with a 40% death rate (27/67 patients, 13 treated with immunotherapy + chemotherapy and 14 treated with immunotherapy only).

Finally, the multimodal score resulting from the combination of clinical, pathomic, and RNA modalities demonstrated a significant association with OS when integrated into a multivariate Cox model along with the clinical features (Figure 6.D). Likelihood-ratio tests indicated a significant effect of this multimodal score compared to a Cox model fitted only with clinical information collected from routine care (p-value = 1.09e-05). Five clinical variables were also significant, including sex, ecog_1, braf, errb2, and pdl1 (see Annex 1). When comparing the multimodal score with the best unimodal scores (i.e., derived from the top performing models for 1-year death prediction, Figure 4.A), only the multimodal score (hazard ratio (HR) = 1.35, 95% CI [1.07-1.69], p-value = 1.11e-02) and the clinical score (HR = 1.25, 95% CI [1.02-1.54], p-value = 3.17e-02) had significantly different performance (Figure 6.D).

Discussion

Multimodal approaches for developing accurate biomarkers for the outcome of metastatic NSCLC patients treated by immunotherapy are highly promising but have been rarely explored so far. In this study, we built a unique multimodal NSCLC cohort to investigate the benefit of integrative strategies. We extracted interpretable features from clinical data, PET/CT images, digitized pathological slides, and bulk RNA-seq profiles, and compared the performance of unimodal and multimodal machine learning models to accurately predict patient outcome. We conducted an extensive exploration of different algorithms, integration strategies, and outcome encodings (i.e., binary vs. continuous targets) to highlight consistent trends that remain robust regardless of the specific choices made within the analysis pipeline.

We trained several unimodal models capable of predicting OS, 1-year death, PFS and 6-months progression using pre-treatment clinical and transcriptomic data. Our design choice to base the entire workflow on interpretable features enabled us to conduct a feature importance analysis. It revealed that clinical models integrated previously established biomarkers into efficient multivariate predictive models (6), while RNA models used signatures of the Tumor Micro-Environment (TME) (20) as well as the expression of specific oncogenes, which were robust to the biopsy location. Notably, our analysis highlighted the positive impact of a high abundance of dendritic cells (DCs), as scored by MCP counter (20), on patient survival. We also found this feature to be weakly associated with favorable prognoses in an RNA-seq cohort of 195 stage III-IV NSCLC patients extracted from The Cancer Genome Atlas (TCGA) (see Methods, Figure s24). These findings thus provide further evidence of the potential of RNA-seq data to predict immunotherapy response. They corroborate recent studies that have demonstrated the enhanced predictive power of RNA-seq data compared to conventional modalities used in clinical practice, such as mutational data or immunohistochemistry (21, 22). Radiomics and pathomics, used as standalone predictors, exhibited limited predictive ability in our analysis. Radiomic models predominantly relied on the TMTV to predict OS and 1-year death, confirming its strong predictive value (23). Nonetheless, the other aggregated features that we investigated did not increase performance, highlighting the need to explore and design additional radiomic features that could effectively complement TMTV.

The importance of DCs for the prediction of patient outcome is in line with previous pre-clinical mechanistic studies that highlighted their central role in shaping anti-tumor immunity (24, 25). In particular, type 1 DCs (DC1) present antigens to CD8+ T cells and their abundance in the TME has been linked to increased survival and improved response to immunotherapy in both animal models (26) and human cancer lesions (27). Tertiary lymphoid structures (TLS) are ectopic formations containing high densities of B cells, T cells, and dendritic cells, at sites of persistent inflammatory stimulation, including tumors (28). Given the accumulating evidence linking the presence of tumor TLS and good prognosis

in cancer patients (29), we investigated a possible association between DCs and TLS. However, visualization of H&E sections, and the poor correlation between B cells and DCs in our data did not support this hypothesis. The link between DCs and prognosis/survival observed in the present study can most likely be explained by DCs' ability to capture tumor antigen in the tumor lesion and present it to T cells in draining lymph nodes.

Our study provided further evidence supporting the superiority of multimodal over unimodal approaches to build accurate biomarkers for the outcome of metastatic NSCLC patients treated with immunotherapy. In all prediction tasks related to OS, 1-year death, PFS, and 6-months progression, we identified multiple multimodal combinations that outperformed the unimodal models, including the models relying on standard clinical data. Furthermore, several combinations demonstrated enhanced patient risk stratification for OS, outperforming the best clinical model across the whole cohort. Multivariate Cox models, combined with Kaplan-Meier analyses, underscored the enhanced prognostic value provided by multimodal predictions beyond routine clinical biomarkers. They further highlighted that multimodal scores could help better identifying patients with the most severe prognosis, thereby guiding tailored treatment strategies such as intensified follow-up care or considering chemotherapy even in cases with high PD-L1 expression. No single integration strategy outperformed others for all prediction tasks, but most of them effectively built multimodal predictors that outperformed the best unimodal predictors. While we confirmed the potential of the DyAM method (16), especially with prior feature selection, we found that a much simpler model based on late fusion frequently compared favorably to more complex integration strategies. We assume that the robust performance of the simple late fusion approach is due to its ability to handle missing modalities. Although it is not ruled out that more complex methods might ultimately yield better results on ideal and large datasets, it should be considered that such ideal scenarios will be quite rare in clinical reality. Overall, our comprehensive benchmark of multiple algorithms and integration strategies highlighted a consistent benefit of combining multiple modalities to predict immunotherapy outcome, irrespective of specific settings or methodologies.

Our study has several limitations. First, we did not have access to an external validation cohort to assess the reproducibility and robustness of our results. We dealt with a relatively modest number of samples for machine learning algorithms, and due to missing modalities, we had to perform our evaluation on 80 patients to ensure a fair comparison between multimodal combinations, thereby limiting the statistical power of our analyses. Therefore, the absolute performance scores highlighted in this study should be interpreted cautiously. Nevertheless, the relative comparison of these scores between different combinations of modalities, as well as in comparison with random predictors, all trained and evaluated with the same methodology, remains valuable. It confirms the superiority of multimodal approaches over unimodal models and thus may motivate the collection of new large and multi-centric multimodal NSCLC cohorts. These multi-centric cohorts are needed to address challenges related to missing modalities, particularly RNA, and further validate our findings. Additionally, our data collection pipeline involved time-consuming processes, such as the manual segmentation and annotation of PET/CT scans by nuclear medicine physicians, which could deter the collection of new external cohorts and limit their size. However, deep learning methods for automatically segmenting lesions on PET scans (30) or computing surrogate radiomic features (31) have recently shown very promising results and may soon be incorporated into the multimodal pipeline to overcome this bottleneck. Finally, despite being one of the most powerful modalities in our analysis, the RNA modality is not yet routinely available in clinical practice in many places, unlike clinical information, PET/CT images, or pathological slides. Its collection involves additional costs and is often affected by the low quality of the remaining tissue samples from the biopsy. However, it could be used as an initial step to identify prognostic mechanisms, which could subsequently be assessed with more cost-effective technologies.

Our unique multimodal cohort allowed us to demonstrate the ability of clinical, radiological, pathological, and transcriptomic data to inform powerful multimodal predictors for the outcome of patients treated with immunotherapy in metastatic NSCLC. It provided several promising predictors that

outperformed both established biomarkers (e.g., PD-L1 expression) and unimodal predictors. They now require refinement and validation in multicentric studies. These results foster further efforts to gather large multimodal cohorts and explore multimodal biomarkers that could efficiently guide therapeutic decision.

Methods

This study was approved by the institutional review board at our institute (DATA200053) and informed consent from all patients was obtained through institutional processes. Data were de-identified, collected and stored in compliance with GDPR.

Clinical data

Baseline clinical data for all 317 patients were collected from the Electronic Medical Record of Institut Curie Hospital using a predefined case report form based on the ESME-AMLC database (NCT03848052).

Each patient's response to immunotherapy was assessed through OS and PFS. OS was defined as the duration from the initiation of first-line immunotherapy (with or without chemotherapy) to the patient's death or last available status update. PFS was defined as the duration from the initiation of first-line immunotherapy to the occurrence of the first progression event or last available status update, including the emergence of new lesions or the progression of pre-existing ones. We also considered binary outcomes, specifically 1-year death (0 for patients who were still alive after one year of immunotherapy and 1 otherwise), and 6-months progression (0 for patients whose disease did not progress after six months of immunotherapy and 1 otherwise). Patients whose OS or PFS was censored before one year or six months, respectively, were excluded from the analysis with binary outcomes.

We selected 30 baseline clinical features for our predictive models. A detailed list of these features and their definitions are provided in Annex 1. We applied one-hot encoding to all categorical features.

Radiomic data

Baseline [18F]FDG-PET /CT scans were collected for 201 patients. Two experienced nuclear medicine physicians delineated all tumor foci in all PET scans using LIFEx software v.7.3 (https://www.lifexsoft.org/) (32). Additionally, they annotated the location of each lesion using an anatomical partition inspired by TNM staging (33). For instance, ipsilateral and contralateral lung metastases were distinguished since they are not associated with the same TNM stage. Subsequently, all images were resampled to a fixed 2x2x2 mm3 voxel size and the segmented tumor regions were processed by applying a fixed threshold of 2.5 standardized uptake value (SUV) units to exclude voxels with SUV values below this threshold. The SUVmax value, the volume, and the centroid of each resulting tumor region were extracted with the IBSI-compliant PyRadiomics Python package (34, 35). The SUVmean values from spherical ROIs manually delineated in the healthy regions of the liver and spleen were also extracted for each patient, without prior 2.5 SUV thresholding (liver ROIs mean volume = 24 cm^3, standard deviation 18 cm^3 – spleen ROIs mean volume = 10 cm^3, standard deviation 8 cm^3).

All these extracted data were then aggregated into 30 baseline whole-body radiomic features to capture the spread of the metastatic disease as well as its metabolic activity. A detailed list of these features and their definitions are provided in Annex 2. We considered the SUVmean values of healthy ROIs in the spleen and the liver, the TMTV (23), and the number of invaded organs visible on the PET scan, including the lungs, sub- and supra-diaphragmatic lymph nodes, the pleura, the liver, the bones, the adrenal gland, and a final category for other regions. Additionally, using the centroids of the processed tumor regions, we calculated the standardized Dmax (36) – the largest distance between two lesions normalized by the body surface area - and the quartile dispersion of the distances between each tumor region's centroid and the global centroid. Finally, for each TNM stage, we took into account all the tumors located in associated regions and computed the TMTV as well as the mean, standard deviation, and maximum value of all the SUVmax values. For instance, for the T stage, we considered the primary

lung tumor as well as all the ipsilateral lung metastases. In cases where no tumor was present in these regions, the feature values were set to zero. Except for features associated with SUVmax, we excluded lesions corresponding to lymphangitic spread, diffuse pleural metastases, diffuse myocardial metastases, and diffuse subdiaphragmatic metastases because their accurate segmentation was questionable.
All the features associated with metabolic volume were log-transformed (i.e., $\log(x+1)$) to deal with right-skewed distributions.

Pathomic data
Baseline pathological slides stained with Hematoxylin-Erythrosine-Saffron (HES) were collected from the FFPE biopsy blocks of 236 patients and subsequently scanned by the Experimental Pathology Platform at Institut Curie. From each slide, we segmented all nuclei with a custom automatic pipeline derived from Lerousseau *et al.* (37), trained in a weakly supervised setting with publicly available data from TCGA as well as data sets from Institut Curie. The slides from the current study were not used to train the segmentation pipeline.
The cell nuclei of six cell types were annotated on each slide, including stromal, epithelial, dead, tumor, connective, and inflammatory cells. These annotations were then used in a pathomic approach (38) to extract 134 relevant features, characterizing the density, the relative proportion, or the spatial organization of these different cell types within the scanned tissue.

Transcriptomic data
Residual FFPE biopsy specimens containing sufficient RNA were collected for 134 patients and RNA sequencing was performed at the Sequencing Core facility of Institut Curie with the Illumina TruSeq RNA Access technology. The RNA-seq data were then processed with the Institut Curie RNA-seq pipeline v4.0.0 (39). The raw bulk RNA-seq read counts were normalized with TPM (Transcripts Per Million) and log-transformed (i.e., $\log(x+1)$).
The abundance of 8 immune cells and 2 stromal cells in the Tumor Micro-Environment was estimated using the MCP-counter method (20). Additionally, log expressions of 22 oncogenes associated with lung cancer were used as features (KRAS, NRAS, EGFR, MET, BRAF, ROS1, ALK, ERBB2, ERBB4, FGFR1, FGFR2, FGFR3, NTRK1, NTRK2, NTRK3, LTK, RET, RIT1, MAP2K1, DDR2, ALK, and CD274). The biopsy site was also considered as a categorical feature and one-hot encoded, distinguishing between lungs, pleura, lymph nodes, bones, liver, adrenal gland, and brain. This information was available for 84 patients. Finally, we used a custom pipeline derived from Jessen *et al.* (40) to estimate the Tumor Mutational Burden (TMB) from the RNA-seq reads mapped to the reference genome hg19 with STAR aligner (1-pass). VarDict tool (41) was used for variant calling and several filters were applied to detect somatic variants. This feature, named TMB_RNA, was available for 110 patients.

Genomic data
TMB was estimated for 42 patients using a custom NGS panel called DRAGON (for Detection of Relevant Alterations in Genes involved in Oncogenetics by NGS) and marketed by Agilent under the name of SureSelect CD Curie CGP. A threshold of 15 mutations per megabase is associated with this panel to distinguish between patients with high TMB and patients with low TMB.

Unimodal analyses - Tree ensemble methods
The Extreme Gradient-Boosted Trees algorithm (19) implemented in XGBoost v.1.7.6 Python package was used to solve 1-year death and 6-months progression classification tasks. We used default parameter values except for *scale_pos_weight* which controls the balance between positive and negative weights and was set to the proportion of negative over positive labels estimated with the training set. Each XGBoost classifier was calibrated with Platt's logistic model (42): predictions were collected with a 10-fold stratified cross-validation scheme on the training set and then used as input for a univariate logistic regression model with balanced class weights. To generate the final "calibrated" predictions, this logistic model was applied to the raw predictions of the XGBoost classifier.

The Random Survival Forest algorithm (18) implemented in scikit-survival v.0.21.0 Python package was used to solve survival tasks for OS and PFS. We used default parameter values except for *max_depth* which controls the size of the survival trees and was set to 6 to mitigate the risk of overfitting. Contrary to XGBoost, Random Survival Forest algorithm does not handle missing values automatically. Therefore, we applied median imputation for continuous features and most-frequent imputation for categorical features, both fitted to the training set.

Unimodal analyses - Linear methods

The Logistic Regression algorithm with elastic net penalty implemented in scikit-learn v.1.2.2 Python package was used to solve 1-year death and 6-months progression classification tasks. We used the saga optimizer, a regularization parameter C=0.1, a L1 ratio of 0.5, a maximum number of iterations of 2500, and balanced class weights.

The Cox's proportional hazard's algorithm with elastic net penalty implemented in scikit-survival v.0.21.0 Python package was used to solve survival tasks for OS and PFS. We used default parameter values except for *alpha_min_ratio* which controls the regularization strength and was set to 0.01.

In both algorithms, we preprocessed the data by first applying robust scaling, followed by median imputation for continuous features and most-frequent imputation for categorical features. All these operations were fitted to the training set.

Multimodal analyses - Late fusion

We used a late fusion strategy to combine every possible subset of modalities. This analysis was limited to fusions of the same predictive algorithms with the same parameter values. For instance, for classification tasks, we separately explored the fusion of penalized logistic regression models and the fusion of XGBoost models.

First, we restricted the training set to patients with at least one of the modalities of the combination of interest available. Then, we independently trained each unimodal model using the subset of patients in the training set for whom the associated modality was available. Finally, for each patient in the test set, we computed the multimodal prediction by averaging the unimodal predictions for the available modalities. For survival models, the unimodal predictions were standardized before averaging, using the mean and standard deviation values estimated for each modality based on predictions obtained from a 10-fold cross-validation scheme applied to the training set (stratified with respect to the censorship rate). XGBoost late fusion models were also calibrated using Platt's logistic model (42), following the strategy described previously.

Multimodal analyses - Early fusion

We used an early fusion strategy to combine every possible subset of data modalities and compared the results with those obtained with the late fusion strategy. The training set was once again limited to patients with at least one available modality. We first pre-processed each data modality separately, considering the subset of patients within the training set for whom that modality was available. Subsequently, we concatenated the processed features from all the modalities to form the input for the predictive model. We used the same models and the same parameter values as in the unimodal analyses (i.e., XGBoost, Logistic Regression, Random Survival Forest, and Cox model). XGBoost early fusion models were calibrated using Platt's logistic model (42).

For linear models, missing modalities were handled by replacing them with zero values. XGBoost did not require special handling for missing modalities, while for Random Survival Forest, a double-coding strategy was applied, inspired by Engemman *et al.* (43). This approach involved duplicating features, assigning either very high values or very low values to patients with missing modalities. This allowed the survival tree to decide on which side of the decision split to place patients with missing modalities.

We also explored early fusion with a preliminary feature selection step to maintain a consistent number of features across different multimodal combinations. We first calculated a univariate score for each feature using $|m - 0.5|$, where m corresponds to either the AUC or the C-index computed with the training set. We then ranked all the features from all the modalities accordingly. To reduce redundancy,

we filtered out highly correlated features by iterating through the ranked feature list from top to bottom and removing subsequent features with a Pearson correlation exceeding $\rho = 0.7$. Finally, for each modality we selected the top $\lfloor n_{total}/n_{modas} \rfloor$ features from the filtered ranked list which belonged to that modality, where n_{total} corresponds to the number of features to keep in the multimodal model (in this analysis we used $n_{total} = 40$) and n_{modas} corresponds to the number of modalities in the multimodal combination of interest. For unimodal combinations this feature selection step was ignored.

Multimodal analyses - DyAM model

We used our own implementation of the DyAM model, with PyTorch v.2.0.1 Python package, to combine every possible subset of data modalities and compared the results with those obtained with late fusion and early fusion strategies. We adopted the exact same architecture as described in Vanguri *et al.* (16), which used single-layer feedforward neural networks with a *tanh* activation function for unimodal predictions and single-layer feedforward neural networks with a *softplus* activation function for unimodal attention weights. Furthermore, similar to (16), we trained our models with a binary cross-entropy loss with balanced class weights, a learning rate of 0.01, 125 training epochs, a L2 regularization strength of 0.001, and the Adam optimizer.

Data were pre-processed with robust scaling as well as median imputation for continuous features and most-frequent imputation for categorical features. For pathomic data we also applied a Principal Component Analysis (PCA) step with 40 components to reduce the size of the neural networks. We also applied a preliminary feature selection step as described previously.

We implemented a nested cross-validation scheme with inner 10-fold stratified cross-validation and a grid-search strategy to optimize the learning rate and the L2 regularization strength for each combination. Due to computational constraints, we limited the number of repetitions to 10 in these cases.

Statistical analysis – performance evaluation

All the predictive models were trained and tested using a 10-fold cross-validation scheme applied to the entire cohort. The folds were stratified based on class proportion for classification tasks and censorship rate for survival tasks. In each fold and for each modality combination, patients with all modalities missing were excluded from the training set. Pre-processing operations, including missing value imputation, scaling, or univariate feature selection, were fitted to each training set, and then applied to the corresponding test set to prevent any data leakage (Figure s4). This process was repeated 100 times, with data shuffling in each iteration. We used the same repeats across all the experiments.

The performance of each model was evaluated using Uno's Concordance Index (C-index) (44) for survival tasks and the Area Under the ROC Curve (AUC) for classification tasks. These metrics were computed for each cross-validation scheme, considering only the predictions in the test sets of the 80 patients with a complete multimodal profile to ensure a fair comparison among the different combinations (Figure s4). Subsequently, the metrics were averaged over all 100 repetitions, and their standard deviation was calculated to measure the variability resulting from the random partition of the data into 10 folds.

Statistical analyses - permutation tests

The significance of the results was assessed with one-sided permutation tests, running the pipeline described above 100 times with randomly shuffled outcomes. Permutation tests were also used for univariate predictors, along with 10,000-repeated bootstrap sampling for computing 95% confidence intervals.

To compare the performance of the different multimodal models and test for statistically significant differences, we applied a two-step procedure. First, for each pair of combinations (i, j) and each cross-validation scheme s, the superiority of j over i was assessed with a one-sided paired permutation test (45), resulting in 100 p-values $\left(p_{ij}^{S}\right)_{i=1}^{100}$. Subsequently, for each pair of combinations (i, j), these 100 p-values were adjusted using the Benjamini-Hochberg procedure (FDR controlled at level α=0.05), and the frequency of statistically significant tests across the 100 tests was computed. The performance of

the different predictive models was estimated on the subset of patients with a complete multimodal profile. Although the paired permutation test described in Bandos *et al.* (45) was originally designed to compare two AUCs, we extended it to the comparison of two C-indexes since both metrics are C statistics and the C-index can be seen as a generalization of the AUC for censored survival data.

Survival analyses

For overall survival, we evaluated the ability of each predictive model to stratify patients into two distinct risk groups, including the models trained to predict PFS or 6-months progression. First, for each fold of each cross-validation scheme, we explored a range of thresholds going from the 30th to the 70th percentiles of the training predictions and selected the one that minimized the log-rank p-value. For PFS-related models, we focused on the stratification of PFS to find the best threshold, mimicking scenarios where OS is not available during the training process. For classification tasks, the 0.5 threshold was also considered. The learned threshold was then applied to the corresponding test set, assigning patients to a low-risk group or a high-risk group for overall survival. Risk group membership was thus collected for each patient across the test sets of the cross-validation scheme. This resulted in 100 group memberships for each patient, corresponding to the 100 cross-validation schemes. Finally, these 100 assignments were aggregated by calculating the frequency of low-risk and high-risk group assignments for each patient. Patients with a frequency of high-risk group greater than 50% were assigned to the final high-risk group, while those with a frequency strictly lower than 50% were assigned to the final low-risk group. We compared the survival distributions of the final low- and high-risk groups for each predictive model with Kaplan-Meier curves and a log-rank test. The Benjamini-Hochberg procedure was used to control for multiple testing (FDR controlled at level $\alpha = 0.05$). We focused on the subset of multimodal combinations which included clinical data to work with a sufficiently large cohort (i.e., 265 patients with the 4 targets available for fair comparisons). This analysis thus assessed the risk stratification ability of multimodal models that incorporated multiple modalities alongside clinical data, whenever they were available.

Finally, we derived a score from each multimodal predictive model by collecting its predictions from the test sets of each cross-validation scheme and averaging them over 100 repeats. This score was used as input to a multivariate Cox model to predict patients' OS, along with clinical features or unimodal scores. The Cox model was then fitted on the 265 patients with the 4 targets available (i.e., OS, 1-year death, PFS, and 6-months progression) for fair comparison between the models. All the input variables were first standardized to ensure comparable hazard ratios. To address missing clinical values, we used median imputation for continuous clinical features and most-frequent imputation for categorical clinical features. For unimodal scores, missing values were replaced by 0.5 for classification models and 0 for survival models, since all the models were "calibrated" with nested cross-validation. We used lifelines v.0.27.4 Python package to fit the Cox models. Likelihood-ratio tests were computed manually with the difference between, the log-likelihood of the two compared models and a chi-squared test.

Feature importance analysis

For each algorithm and each modality, we used the permutation explainer provided by SHAP v0.42.1 Python package (46) to compute the SHAP values for each feature and each patient. SHAP values were computed only when the patient was in a test set of the cross-validation scheme. This resulted in $n_{patients} \times n_{features}$ SHAP values for each cross-validation scheme, where $n_{patients}$ corresponds to the number of patients with the modality of interest available and $n_{features}$ corresponds to the number of features extracted for this modality. All these values were subsequently averaged across the 100 cross-validation schemes to produce the final set of $n_{patients} \times n_{features}$ mean SHAP values.

A positive SHAP value for a patient p and a feature f means that considering the feature f in the predictive model of interest increases the patient p's probability of death or progression for classification tasks and the patient p's risk of death or progression for survival tasks. Conversely, a negative SHAP value means that f decreases patient p's probability or risk of death or progression.

For each data modality, we applied a three steps procedure to combine the SHAP values from both related tasks (i.e., OS and 1-year death, or PFS and 6-months progression) and both approaches (i.e., linear and tree ensemble methods) and obtain a consensus ranking of features with respect to their importance in the predictive models for overall survival or progression-free survival. First, we filtered out non-robust features whose impact on the predictions was not consistent across the four predictive models (for radiomics we did not consider the LR model since its AUC was lower than 0.5). To do so we computed, for each feature and each model, the Spearman correlation between the SHAP values and the values of the feature of interest and then filtered out the features whose correlation sign was not consistent across the four models. Then we ranked the remaining robust features for each of the four models with respect to their absolute SHAP values averaged across all the patients. For each model $i \in \{1, \dots, 4\}$ and each feature f we thus obtained a rank $r_f^i \in \{1, \dots, n_F\}$ (with n_F the number of robust features) with 1 corresponding to the least important feature and n_F to the most important one. Finally, we aggregated all these ranks across the four models to obtain a consensus ranking, taking into account the performance of each model. The consensus rank of feature f was defined as:

$$r_f^{cons} = \frac{1}{s_1 + s_2 + s_3 + s_4} \sum_{i=1}^{4} s_i \, r_i^f \tag{1}$$

Where s_i corresponds to the score of the model i and is equal to $\max(0, score_i - 0.5)$ (the score is either the AUC or the C-index). The remaining robust features were also tested with univariate permutation tests both for OS and 1-year death (or PFS and 6-months progression): 1000 univariate AUCs or C-indexes were generated with permutated labels and then compared to the original AUC or C-index. Features that remained statistically significant after Benjamini-Hochberg correction (FDR controlled at the level α=0.05) were reported in the consensus ranking.
The consensus ranks were normalized with respect to the total number of consensus features. Each rank was also assigned a sign which corresponded to the sign of the Spearman correlation coefficient between the SHAP values and the values of the associated feature. A positive sign means that the effect of the feature on the predicted risk/probability of event increases with the feature value, while a negative sign means that the effect decreases with the feature value. In this context, the term "effect" is linked to the SHAP value and can be either positive (increasing predicted risk) or negative (decreasing predicted risk).

TCGA analysis
Bulk RNA-seq profiles of the primary tumor of 518 NSCLC patients with adenocarcinoma (TCGA-LUAD) and 504 patients with squamous cell carcinoma (TCG-LUSC) were extracted with the GDC data portal (47) and normalized with TPM. Clinical data and overall survival information for these patients were gathered from the study conducted by Liu *et al.* (48). Unfortunately, detailed treatment information was not available for most of the patients.
Patients diagnosed at an early stage were filtered out, resulting in a sub-cohort of 195 patients with stage III-IV NSCLC. We then calculated the same transcriptomic features as previously described (i.e., MCP counter signatures and log-expression of specific oncogenes). The biopsy site was not considered since all the samples originated from the primary lung tumor, and we did not compute the TMB_RNA. We used the same algorithms as described above to predict OS and 1-year death. Models were trained and tested with stratified 10-fold cross-validation schemes repeated 100 times. Finally, the previously described feature importance analysis was performed to highlight consensus important features.

Data Availability

The raw data generated in this study, including PET/CT scans, digitized pathological slides, and RNA-seq profiles, are not publicly available due to patient privacy requirements. Derived transcriptomic, radiomic, and pathomic features as well as clinical information are available upon reasonable request from the corresponding author.

The transcriptomic features as well as the clinical features associated with the 195 TCGA patients with stage III-IV NSCLC are publicly available (https://github.com/ncaptier/multipit/tree/main/data).

Code Availability

We have made all our codes available in a GitHub repository with associated documentation allowing for the reproduction of our multimodal analyses with external data, additional modalities, or different features (https://github.com/ncaptier/multipit).

Acknowledgements

We would like to thank the following collaborators at Institut Curie for their valuable support in data management and processing: A. Nicolas, R. Goudefroye, C. Martinat, M. Bouvet, and A. Vincent-Salomon from the experimental pathology platform, L. Chanas, and M. Milder from Institut Curie's Data Office, T. Ramtohul, and H. Brisse from the Department of Radiology, S. Baulande from the Next-Generation Sequencing platform, I. Bièche, and C. Callens from the Diagnostic and Theranostic Medicine Division, C. Reyes, A. Rapinat, and D. Gentien from the Genomics platform, and C. Kamoun, N. Servant, and P. Hupé from the bioinformatics core facility. We also thank M. Lefevre and S. Lefranc from Institut Mutualiste Montsouris.

This work was part of the TIPIT project (Towards an Integrative approach for Precision ImmunoTherapy) funded by Fondation ARC call «SIGN'IT 2020—Signatures in Immunotherapy». The present study was also supported by the French government under management of Agence Nationale de la Recherche as part of the 'Investissements d'avenir' program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

1. Hendriks LE, Kerr KM, Menis J, Mok TS, Nestle U, Passaro A, *et al*. Non-oncogene-addicted metastatic non-small-cell lung cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. Ann Oncol. 2023 Apr;34(4):358-376.
2. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, *et al*. Updated Analysis of KEYNOTE-024: Pembrolizumab Versus Platinum-Based Chemotherapy for Advanced Non-Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score of 50% or Greater. J Clin Oncol. 2019 Mar 1;37(7):537-546.
3. Gandhi L, Rodríguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, *et al*. Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. N Engl J Med. 2018 May 31;378(22):2078-2092.
4. Paz-Ares L, Vicente D, Tafreshi A, Robinson A, Soto Parra H, Mazières J, *et al*. A Randomized, Placebo-Controlled Trial of Pembrolizumab Plus Chemotherapy in Patients With Metastatic Squamous NSCLC: Protocol-Specified Final Analysis of KEYNOTE-407. J Thorac Oncol. 2020 Oct;15(10):1657-1669.
5. Herbst RS, Giaccone G, de Marinis F, Reinmuth N, Vergnenegre A, Barrios CH, *et al*. Atezolizumab for First-Line Treatment of PD-L1-Selected Patients with NSCLC. N Engl J Med. 2020 Oct 1;383(14):1328-1339.
6. Hellmann MD, Paz-Ares L, Bernabe Caro R, Zurawski B, Kim SW, Carcereny Costa E, *et al*. Nivolumab plus Ipilimumab in Advanced Non-Small-Cell Lung Cancer. N Engl J Med. 2019 Nov 21;381(21):2020-2031.
7. Joshi I, Peravali M, Geng X, Rao S, Chen KY, Veytsman I, *et al*. Impact of Baseline Clinical Biomarkers on Treatment Outcomes in Patients With Advanced NSCLC Receiving First-line Pembrolizumab-Based Therapy. Clin Lung Cancer. 2022 Jul;23(5):438-445.

8. Ahn BC, So JW, Synn CB, Kim TH, Kim JH, Byeon Y, *et al*. Clinical decision support algorithm based on machine learning to assess the clinical response to anti-programmed death-1 therapy in patients with non-small-cell lung cancer. Eur J Cancer. 2021 Aug;153:179-189.
9. Trebeschi S, Drago SG, Birkbak NJ, Kurilova I, Călin AM, Delli Pizzi A, *et al*. Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers. Ann Oncol. 2019 Jun 1;30(6):998-1004.
10. Mu W, Tunali I, Gray JE, Qi J, Schabath MB, Gillies RJ. Radiomics of 18F-FDG PET/CT images predicts clinical benefit of advanced NSCLC patients to checkpoint blockade immunotherapy. Eur J Nucl Med Mol Imaging. 2020 May;47(5):1168-1182.
11. Deng J, Yang J, Hou L, Wu J, He Y, Zhao M, *et al*. Genopathomic profiling identifies signatures for immunotherapy response of lung adenocarcinoma via confounder-aware representation learning. iScience. 2022 Oct 17;25(11):105382.
12. Barrera C, Corredor G, Viswanathan VS, Ding R, Toro P, Fu P, *et al*. Deep computational image analysis of immune cell niches reveals treatment-specific outcome associations in lung cancer. NPJ Precis Oncol. 2023 Jun 1;7(1):52.
13. Patil NS, Nabet BY, Müller S, Koeppen H, Zou W, Giltnane J, *et al*. Intratumoral plasma cells predict outcomes to PD-L1 blockade in non-small cell lung cancer. Cancer Cell. 2022 Mar 14;40(3):289-300.e4.
14. Ravi A, Hellmann MD, Arniella MB, Holton M, Freeman SS, Naranbhai V, *et al*. Genomic and transcriptomic analysis of checkpoint blockade response in advanced non-small cell lung cancer. Nat Genet. 2023 May;55(5):807-819.
15. Zhao Q, Xie R, Lin S, You X, Weng X. Anti-PD-1/PD-L1 Antibody Therapy for Pretreated Advanced or Metastatic Nonsmall Cell Lung Carcinomas and the Correlation between PD-L1 Expression and Treatment Effectiveness: An Update Meta-Analysis of Randomized Clinical Trials. Biomed Res Int. 2018 Sep 24;2018:3820956.
16. Vanguri RS, Luo J, Aukerman AT, Egger JV, Fong CJ, Horvat N, *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. Nat Cancer. 2022 Oct;3(10):1151-1164.
17. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on typical tabular data?. Advances in Neural Information Processing Systems. 2022 Dec 6;35:507-20.
18. Ishwaran H, Kogalur U, Blackstone E, Lauer M. Random survival forests. The Annals of Applied Statistics, 2(3) 841-860 September 2008.
19. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16).
20. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, *et al*. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016 Oct 20;17(1):218.
21. Sun D, Liu J, Zhou H, Shi M, Sun J, Zhao S, *et al*. Classification of Tumor Immune Microenvironment According to Programmed Death-Ligand 1 Expression and Immune Infiltration Predicts Response to Immunotherapy Plus Chemotherapy in Advanced Patients With NSCLC. J Thorac Oncol. 2023 Jul;18(7):869-881.
22. Li A, Wang Y, Yu Z, Tan Z, He L, Fu S, *et al*. STK11/LKB1-Deficient Phenotype Rather Than Mutation Diminishes Immunotherapy Efficacy and Represents STING/Type I Interferon/CD8+ T-Cell Dysfunction in NSCLC. J Thorac Oncol. 2023 Dec;18(12):1714-1730.
23. Seban RD, Mezquita L, Berenbaum A, Dercle L, Botticella A, Le Pechoux C, *et al*. Baseline metabolic tumor burden on FDG PET/CT scans predicts outcome in advanced NSCLC patients treated with immune checkpoint inhibitors. Eur J Nucl Med Mol Imaging. 2020 May;47(5):1147-1157.

24. Broz ML, Binnewies M, Boldajipour B, Nelson AE, Pollack JL, Erle DJ, *et al*. Dissecting the tumor myeloid compartment reveals rare activating antigen-presenting cells critical for T cell immunity. Cancer Cell. 2014 Nov 10;26(5):638-52.
25. Sánchez-Paulete AR, Cueto FJ, Martínez-López M, Labiano S, Morales-Kastresana A, Rodríguez-Ruiz ME, *et al*. Cancer Immunotherapy with Immunomodulatory Anti-CD137 and Anti-PD-1 Monoclonal Antibodies Requires BATF3-Dependent Dendritic Cells. Cancer Discov. 2016 Jan;6(1):71-9.
26. Salmon H, Idoyaga J, Rahman A, Leboeuf M, Remark R, Jordan S, *et al*. Expansion and Activation of CD103(+) Dendritic Cell Progenitors at the Tumor Site Enhances Tumor Responses to Therapeutic PD-L1 and BRAF Inhibition. Immunity. 2016 Apr 19;44(4):924-38.
27. Barry KC, Hsu J, Broz ML, Cueto FJ, Binnewies M, Combes AJ, *et al*. A natural killer-dendritic cell axis defines checkpoint therapy-responsive tumor microenvironments. Nat Med. 2018 Aug;24(8):1178-1191.
28. Goc J, Germain C, Vo-Bourgais TK, Lupo A, Klein C, Knockaert S, *et al*. Dendritic cells in tumor-associated tertiary lymphoid structures signal a Th1 cytotoxic immune contexture and license the positive prognostic value of infiltrating CD8+ T cells. Cancer Res. 2014 Feb 1;74(3):705-15.
29. Cabrita R, Lauss M, Sanna A, Donia M, Skaarup Larsen M, Mitra S, *et al*. Tertiary lymphoid structures improve immunotherapy and survival in melanoma. Nature. 2020 Jan;577(7791):561-565.
30. Park J, Kang SK, Hwang D, Choi H, Ha S, Seo JM, *et al*. Automatic Lung Cancer Segmentation in [18F]FDG PET/CT Using a Two-Stage Deep Learning Approach. Nucl Med Mol Imaging. 2023 Apr;57(2):86-93.
31. Girum KB, Rebaud L, Cottereau AS, Meignan M, Clerc J, Vercellino L, *et al*. 18F-FDG PET Maximum-Intensity Projections and Artificial Intelligence: A Win-Win Combination to Easily Measure Prognostic Biomarkers in DLBCL Patients. J Nucl Med. 2022 Dec;63(12):1925-1932.
32. Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, *et al*. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. Cancer Res. 2018 Aug 15;78(16):4786-4789.
33. Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, *et al*. The IASLC Lung Cancer Staging Project: Proposals for Revision of the TNM Stage Groupings in the Forthcoming (Eighth) Edition of the TNM Classification for Lung Cancer. J Thorac Oncol. 2016 Jan;11(1):39-51.
34. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, *et al*. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020 May;295(2):328-338.
35. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, *et al*. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. 2017 Nov 1;77(21):e104-e107.
36. Cottereau AS, Meignan M, Nioche C, Capobianco N, Clerc J, Chartier L, *et al*. Risk stratification in diffuse large B-cell lymphoma using lesion dissemination and metabolic tumor burden calculated from baseline PET/CT†. Ann Oncol. 2021 Mar;32(3):404-411.
37. Lerousseau M, Vakalopoulou M, Classe M, Adam J, Battistella E, Carré A, *et al*. Weakly Supervised Multiple Instance Learning Histopathological Tumor Segmentation. MICCAI 2020 - Medical Image Computing and Computer Assisted Intervention, Oct 2020, Lima, Peru.
38. Bülow RD, Hölscher DL, Costa IG, Boor P. Extending the landscape of omics technologies by pathomics. NPJ Syst Biol Appl. 2023 Aug 7;9(1):38.
39. Servant N, La Rosa P, phupe, Allain F. (2023). bioinfo-pf-curie/RNA-seq: v4.0.0 (v4.0.0). Zenodo. https://doi.org/10.5281/zenodo.7837455
40. Jessen E, Liu Y, Davila J, Kocher JP, Wang C. Determining mutational burden and signature using RNA-seq from tumor-only samples. BMC Med Genomics. 2021 Mar 1;14(1):65.

41. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, *et al*. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016 Jun 20;44(11):e108.
42. Niculescu-Mizil A, Caruana R. (2005). Predicting good probabilities with supervised learning. ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning. 625-632. 10.1145/1102351.1102430.
43. Engemann DA, Kozynets O, Sabbagh D, Lemaître G, Varoquaux G, Liem F, *et al*. Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. Elife. 2020 May 19;9:e54055.
44. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011 May 10;30(10):1105-17.
45. Bandos AI, Rockette HE, Gur D. A permutation test sensitive to differences in areas for comparing ROC curves from a paired design. Stat Med. 2005 Sep 30;24(18):2873-93.
46. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30
47. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016 Sep 22;375(12):1109-12.
48. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, *et al*. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018 Apr 5;173(2):400-416.e11.

Figures

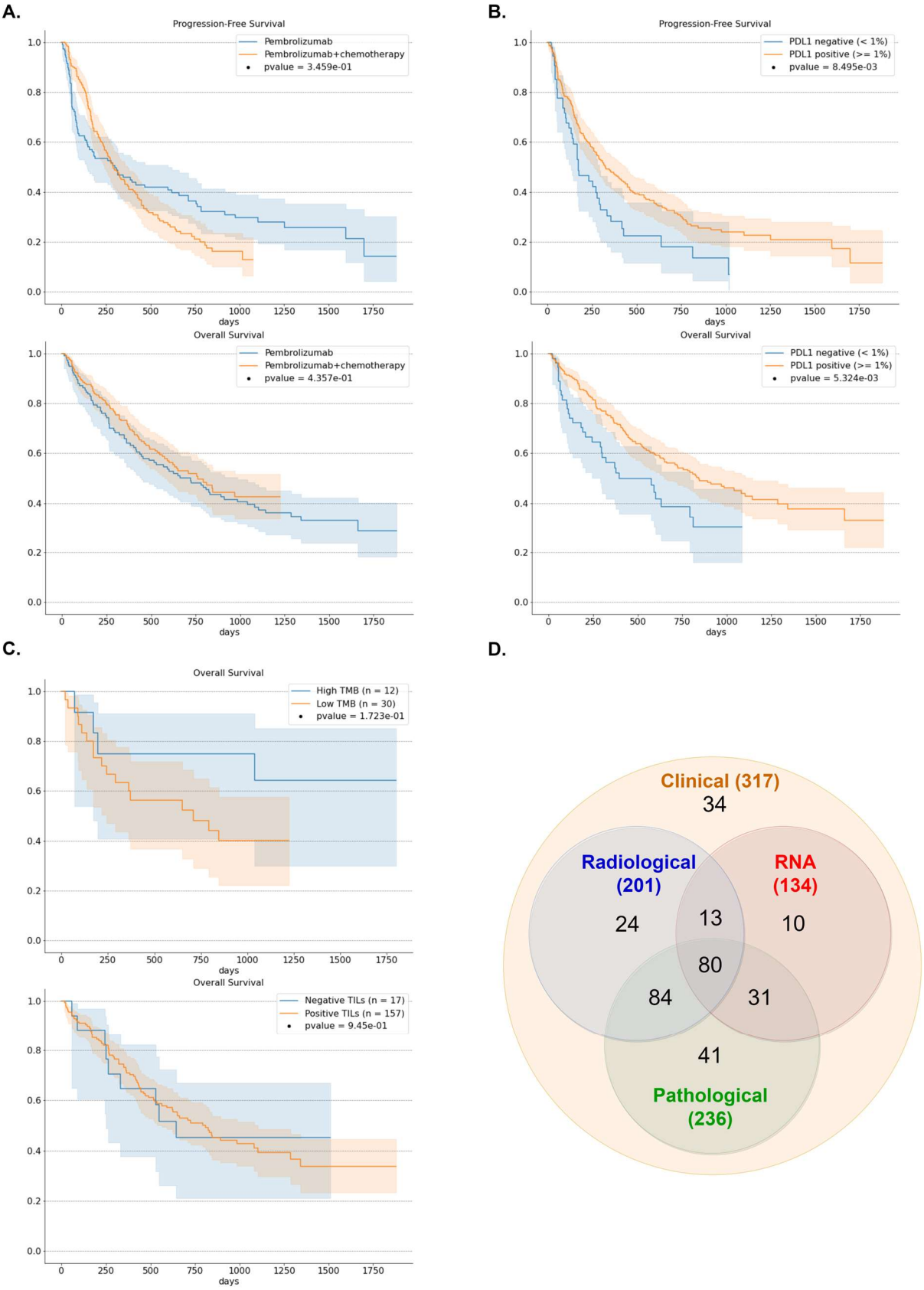
A.
Progression-Free Survival
Pembrolizumab
Pembrolizumab+chemotherapy
pvalue = 3.459e-01
days
Overall Survival
Pembrolizumab
Pembrolizumab+chemotherapy
pvalue = 4.357e-01
days
B.
Progression-Free Survival
PDL1 negative (< 1%)
PDL1 positive (>= 1%)
pvalue = 8.495e-03
days
Overall Survival
PDL1 negative (< 1%)
PDL1 positive (>= 1%)
pvalue = 5.324e-03
days
C.
Overall Survival
High TMB (n = 12)
Low TMB (n = 30)
pvalue = 1.723e-01
days
Overall Survival
Negative TILs (n = 17)
Positive TILs (n = 157)
pvalue = 9.45e-01
days
D.
Clinical (317)
34
Radiological
(201)
RNA
(134)
24
13
10
80
84
31
41
Pathological
(236)

Figure 1: A. OS and PFS Kaplan-Meier survival curves for the whole NSCLC cohort (n=311 for OS and n=316 for PFS) with 95% confidence interval. Patients are stratified with respect to their first-line therapy, either pembrolizumab alone or pembrolizumab + chemotherapy. Log-rank p-values are reported to characterize the separation of the survival curves. **B.** OS and PFS Kaplan-Meier survival curves with 95% confidence interval for the patients with available PD-L1 expression (n=295 for OS and n=300 for PFS). Patients are stratified with respect to their PD-L1 status (positive vs negative). **C.** OS Kaplan-Meier survival curves with 95% confidence interval for the 42 patients with available TMB and the 174 patients with available TILs status. For the TMB, patients are stratified with a threshold of 15 mutations per megabase (supplementary methods section 1.e). For TILs, patients are stratified with respect to their positive vs negative TILs status. **D.** Overview of the multimodal cohort with a Venn diagram. The four data modalities and their intersections are represented (i.e., PET/CT images, clinical data, pathological slides, and bulk RNA-seq profiles).

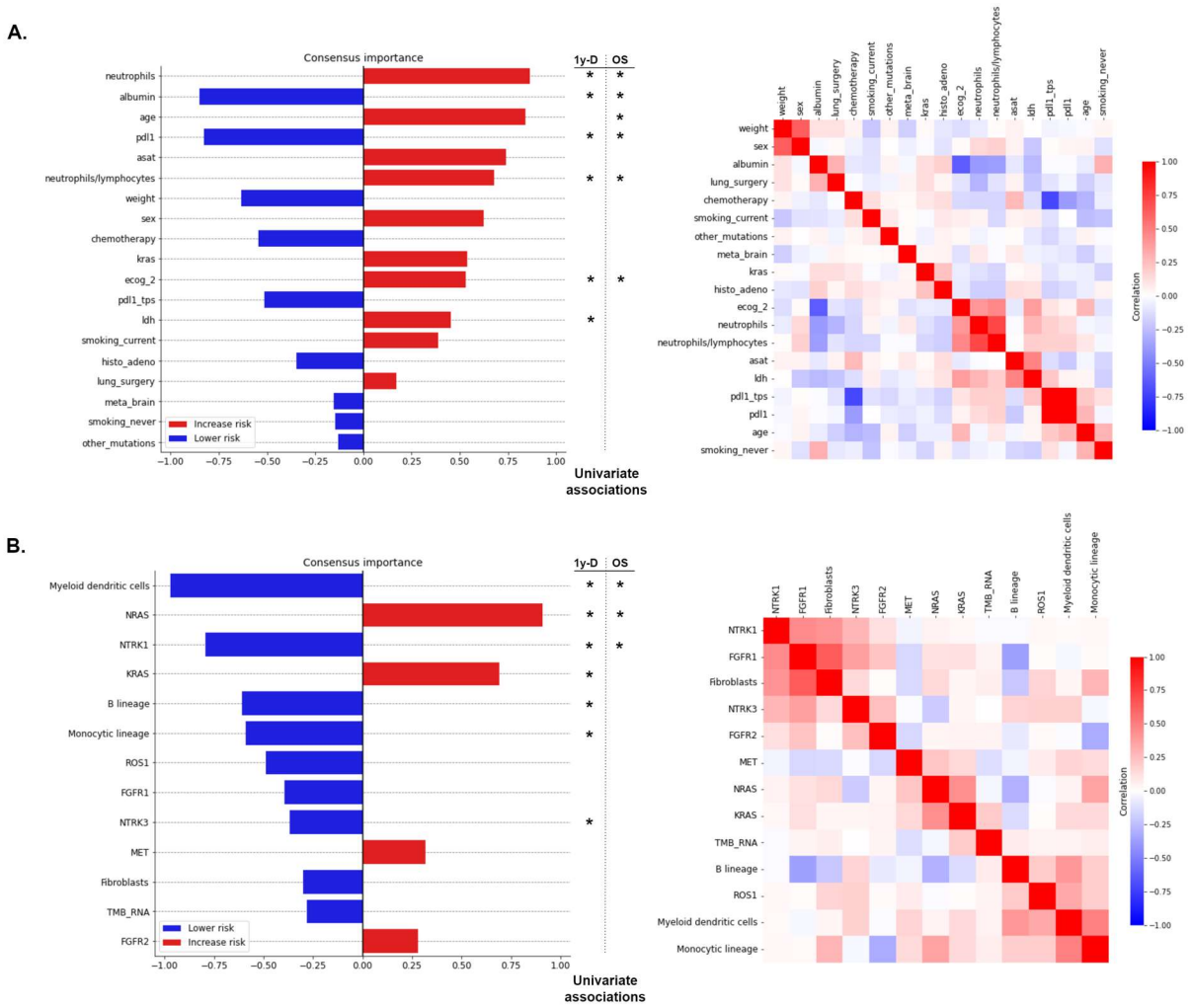

Figure 2: Feature importance ranking for the prediction of overall survival, obtained by aggregating the SHAP values collected from both tasks (OS and 1-year death) and both approaches (linear and tree ensemble methods) (supplementary methods 2.c). Features that are significantly associated with 1-year death (permutation test with univariate AUCs) after Benjamini-Hochberg (BH) correction ($\alpha = 0.05$) are shown with a * on the left side, while features that are significantly associated with OS (permutation test with univariate C-index) after BH correction are annotated with a * on the right side. **A.** Consensus feature importance ranking for the clinical data modality (left) and heatmap of correlations between consensus clinical features (right). Correlations were evaluated by Spearman correlation coefficients (for continuous feature vs continuous feature), AUCs rescaled to $[-1, 1]$ (for continuous feature vs binary categorical feature), or Matthews correlation coefficient (for binary categorical feature vs binary categorical feature). **B.** Consensus feature importance ranking for the RNA data modality (left) and heatmap of Spearman correlations between consensus RNA features (right).

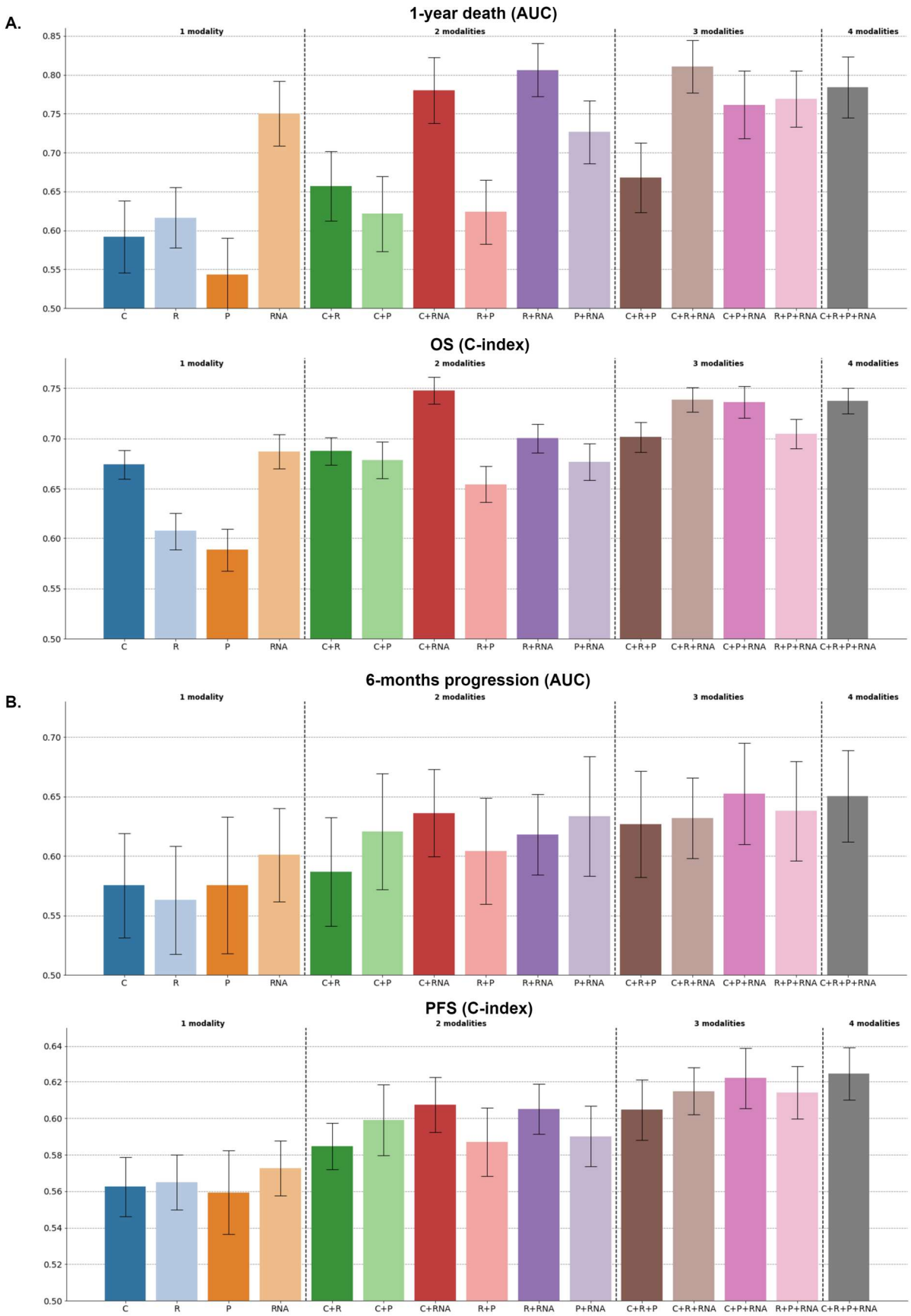

Figure 3: Performance of all the possible multimodal combinations, with a late fusion strategy and tree ensemble methods (C: clinical, R: radiomic, P: pathomic, RNA). The bar height corresponds to the performance metric (either ROC AUC or C-index) averaged across the 100 cross-validation schemes,

and the error bar corresponds to ± 1 standard deviation, estimated across the 100 cross-validation schemes. **A.** ROC AUCs associated with the prediction of 1-year death with XGBoost algorithms (top) and estimated with n=77 patients. C-indexes associated with the prediction of OS with Random Survival Forest algorithms (bottom) and estimated with n=79 patients. **B**. ROC AUCs associated with the prediction of 6-months progression with XGBoost algorithms (top) and estimated with n=75 patients. C-indexes associated with the prediction of PFS with Random Survival Forest algorithms (bottom) and estimated with n=80 patients.

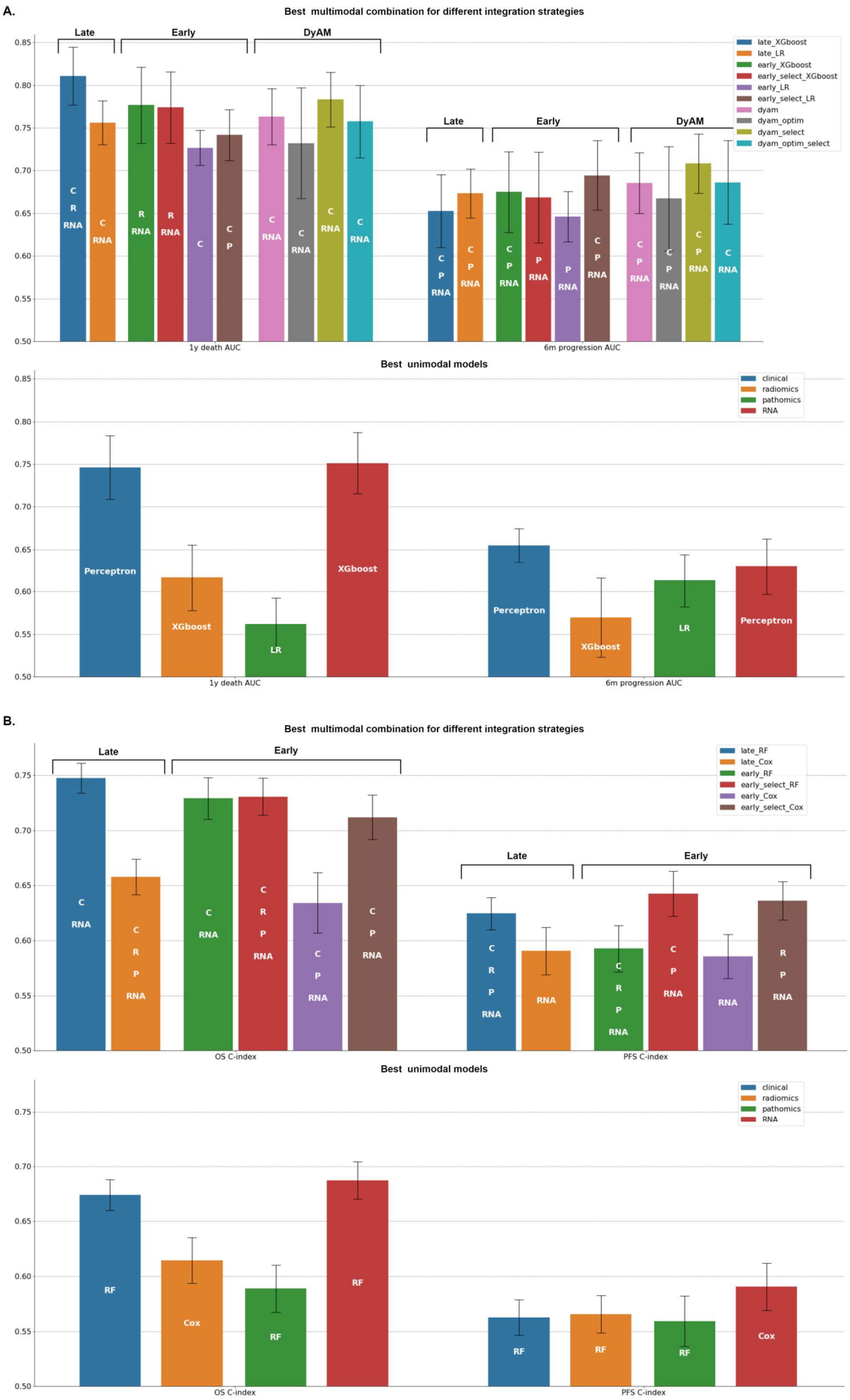
A.
Best multimodal combination for different integration strategies
Late
Early
DyAM
late_XGboost
late_LR
early_XGboost
early_select_XGboost
early_LR
early_select_LR
dyam
dyam_optim
dyam_select
dyam_optim_select
1y death AUC
6m progression AUC
Best unimodal models
clinical
radiomics
pathomics
RNA
Perceptron
XGboost
LR
B.
Best multimodal combination for different integration strategies
late_RF
late_Cox
early_RF
early_select_RF
early_Cox
early_select_Cox
OS C-index
PFS C-index
Best unimodal models
RF
Cox

Figure 4: Best unimodal and multimodal performances across all the possible combinations of modalities and predictive algorithms. The top barplot displays the performance of the best multimodal combination for each integration strategy, while the bottom barplot shows the performance of the best unimodal algorithm for each data modality. Bar heights and error bars correspond to the mean metric (AUC or C-index) and ± 1 standard deviation respectively, estimated across the 100 cross-validation schemes (except for the dyam_optim models for which only 10 cross-validation schemes were used, due to computational constraints). **A.** Best performances (AUC) for the prediction of 1-year death and 6-months progression (n=77 for 1-year OS and n=75 for 6-months PFS). **B.** Best performances (C-index) for the prediction of OS and PFS (n=79 for OS and n=80 for PFS).

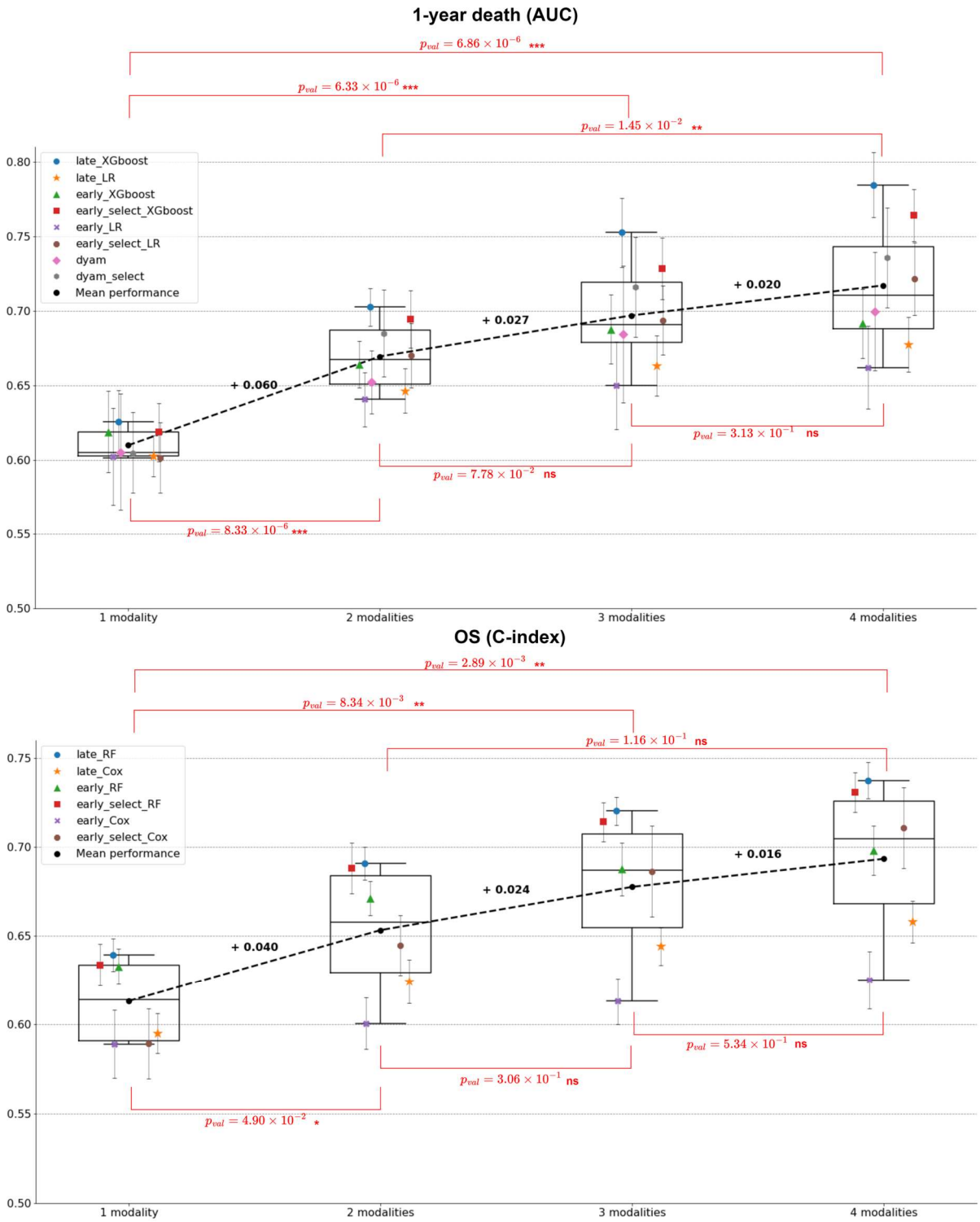

Figure 5: Average performance across all models with 1 modality, 2 modalities, 3 modalities and 4 modalities for 1-year death (top) and OS (bottom). Markers and error bars correspond to the mean average performance and ± 1 standard deviation respectively, estimated across the 100 cross-validation schemes. Mean increases are represented with dashed lines and bold annotations. Red annotations correspond to Student's t test p-values to compare the different numbers of integrated modalities (e.g., 1 modality vs 2 modalities).

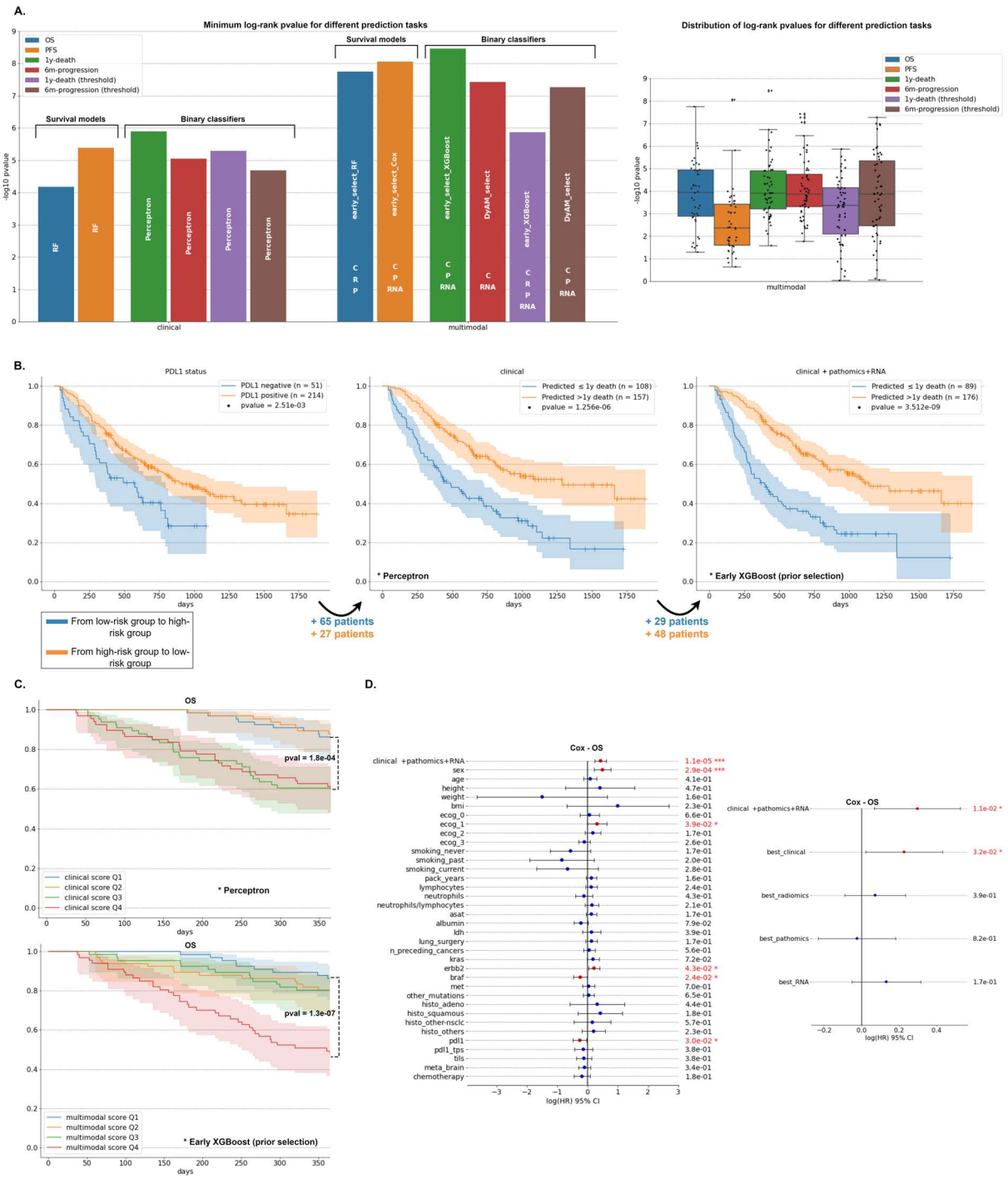

Figure 6: A. Comparison of the stratification of the patients into high-risk and low-risk groups for OS, for different predictive tasks, with Kaplan-Meier survival analysis and log-rank p-values (n=265 patients with the 4 targets available for a fair comparison). Only the combinations including the clinical modality are compared (see supplementary methods section 4.b). **B.** Kaplan-Meier survival curves for the high-risk and low-risk OS groups defined by PDL1-status (left), the clinical model with the lowest log-rank p-value (middle), and the multimodal model with lowest log-rank p-value (right). **C.** Kaplan-Meier survival curves for OS within the first year of therapy. The cohort is stratified into quartiles based on either the clinical score derived from the clinical perceptron predictions (top), or the multimodal score derived from the predictions of the clinical+pathomics+RNA model (bottom). **D.** Hazard ratios and likelihood-ratio test p-values associated with multivariate Cox models trained to predict patient's OS. Cox model with the clinical+pathomics+RNA score as well as the clinical features collected in this study (left). Cox model with the clinical+pathomics+RNA score as well as the best unimodal scores derived

from the top performing unimodal models for 1-year death prediction (right), with the best clinical score corresponding to the clinical perceptron identified in panels B and C.

Tables

Table 1: Clinical characteristics of the multimodal cohort and the subset of patients with a complete multimodal profile. The results of the statistical comparison between the subset of patients with a complete profile and the rest of the cohort, using Welch's t-tests, Chi-squared tests, and Log-rank tests, are presented in the last column.

Clinical characteristics		**Multimodal cohort (n = 317)**	**Immuno + chemo (n=196)**	**Immuno alone (n=121)**	**Subset with all modalities (n=80)**	**Statistical comparison (80 vs 237)**
Age – median (range)		66 (33-92)	64 (33-84)	69 (40-92)	64 (37-82)	p^1_{val} = 4.9e-3
Sex - n (%)	Men	128 (40)	83 (42)	45 (37)	34 (43)	p^2_{val} = 6.0e-1
	Women	189 (60)	113 (58)	76 (63)	46 (57)	
1st line therapy – n (%)	Pembrolizumab + chemotherapy	196 (62)	196 (100)	-	55 (69)	p^2_{val} = 9.2e-2
	Pembrolizumab	121 (38)	-	121 (100)	25 (31)	
Histology - n (%)	Adenocarcinomas	232 (73)	152 (77)	80 (66)	54 (68)	p^2_{val} = 2.8e-2
	Squamous cell carcinomas	44 (14)	17 (9)	27 (22)	13 (16)	
	Other subtypes/not available	41 (13)	27 (14)	14 (12)	13 (16)	
PD-L1 expression – n (%)	≥ 50%	163 (51)	49 (25)	114 (94)	42 (52)	p^2_{val} = 6.2e-1
	1% - 49%	82 (26)	78 (40)	4 (3)	23 (29)	
	Negative	56 (18)	56 (29)	0 (0)	11 (14)	
	Not available	16 (5)	13 (6)	3 (3)	4 (5)	
Smoking status – n (%)	Current/former	287 (91)	180 (92)	107 (88)	71 (89)	p^2_{val} = 6.9e-1
	Never	29 (9)	15 (8)	14 (12)	8 (10)	
	Not available	1 (<1)	1 (<1)	0 (0)	1 (1)	
Performance status – n (%)	ECOG 0/1	244 (77)	158 (81)	86 (71)	71 (89)	p^2_{val} = 1.5e-3
	ECOG ≥2	36 (11)	14 (7)	22 (18)	1 (1)	
	Not available	37 (12)	24 (12)	13 (11)	8 (10)	
TILs – n (%)	Positive	159 (50)	82 (42)	77 (64)	49 (61)	p^2_{val} = 2.8e-2
	Negative	18 (6)	9 (5)	9 (7)	3 (4)	
	Not available	140 (44)	105 (53)	35 (29)	28 (35)	
Median Overall Survival – days (95% CI)		756 (592-910)	723 (446-987)	763 (576-NR)	846 (650-NR)	p^3_{val} = 2.7e-1
Median Progression Free Survival – days (95% CI)		296 (241-372)	301 (145-598)	290 (241-372)	386 (275-711)	p^3_{val} = 1.6e-2

ECOG: Eastern Cooperative Oncology Group
TILs: Tumor-Infiltrating Lymphocytes
NR: Not Reached
p^1_{val}: Welch's t-test p-value - p^2_{val}: Chi-squared test p-value - p^3_{val}: Log-rank test p-value

Table 2: Unimodal performance of each data modality for the prediction of OS, 1-year death, PFS, and 6-months progression with linear and tree ensemble algorithms (mean ± std over the 100 cross-validation schemes). The best performances for each column are highlighted in bold.

Target (number of patients)		**OS (n=79)**	**1-year death (n=77)**	**PFS (n=80)**	**6-months progression (n=75)**
Metric		**C-index**	**AUC**	**C-index**	**AUC**
Clinical	Tree ensembles	0.67±0.01 *	0.59±0.05	0.56±0.02	0.58±0.04
	Linear	0.60±0.02 *	0.73±0.02 *	0.53±0.03	**0.61±0.03** *
Radiomics	Tree ensembles	0.61±0.02 *	0.62±0.04	0.57±0.01	0.56±0.05
	Linear	0.61±0.02 *	0.47±0.03	0.55±0.02	0.48±0.04
Pathomics	Tree ensembles	0.59±0.02	0.54±0.05	0.56±0.02	0.58±0.06 *
	Linear	0.58±0.02	0.56±0.03	0.51±0.02	**0.61±0.03** *
RNA	Tree ensembles	**0.69±0.02** *	**0.75±0.04** *	0.57±0.02	0.60±0.04 *
	Linear	0.58±0.02	0.65±0.03	**0.59±0.02** *	**0.61±0.03**

*: permutation p-value ≤ 0.05 (exact p-values are reported in Table s1)