

Supplementary Materials

For

Radiomic analysis of patient and inter-organ heterogeneity in response to immunotherapies and BRAF targeted therapy in metastatic melanoma

Running Title: melanoma organ-specific radiomics

Feature: Immunotherapy Biomarkers

Authors: Alexandra Tompkins^{1,2*}, Zane N. Gray^{1,2*}, Rebekah E. Dadey^{1,2}, Serafettin Zenkin^{1,3}, Nasim Batavani^{1,3}, Sarah Newman^{1,2}, Afsaneh Amouzegar^{1,2}, Murat Ak^{1,3}, Nursima Ak^{1,3}, Taha Yasin Pak^{1,3}, Vishal Peddagangireddy^{1,3}, Priyadarshini Mamindla^{1,3}, Sarah Behr¹, Amy Goodman¹, Darcy L. Ploucha¹, John M. Kirkwood^{1,2}, Hassane M. Zarour^{1,4}, Yana G. Najjar^{1,2}, Diwakar Davar^{1,2}, Rivka Colen^{1,3}, Jason J. Luke^{1,2#}, Riyue Bao^{1,2#}

¹UPMC Hillman Cancer Center, Pittsburgh, PA

²University of Pittsburgh, Department of Medicine, Pittsburgh, PA

³University of Pittsburgh Department of Radiology, Pittsburgh, PA

⁴University of Pittsburgh, Department of Immunology, Pittsburgh, PA

*Equal Contribution.

#Co-senior and corresponding authors.

#Corresponding Authors:

Jason J. Luke, MD, FACP

Associate Professor of Medicine

UPMC Hillman Cancer Center

5150 Centre Ave. Room 1.27C

Pittsburgh, PA 15232

Telephone: (412) 647-2811

Fax: (412) 623-7948

lukejj@upmc.edu

Riyue Bao, PhD

Associate Professor of Medicine

UPMC Hillman Cancer Center

5150 Centre Ave, Suite 1A, Room 105

Pittsburgh, PA 15232

Telephone: (412) 623-0021

Fax: (412) 864-7603

baor@upmc.edu

Keywords: immunotherapy, checkpoint inhibitors, radiomics, machine learning, response prediction, organ-specific, metastatic, melanoma

This document contains the full description of **Methods** for the manuscript, and **Supplementary Figures 1 to 9. Supplementary Tables 1 to 5** are provided in a separate spreadsheet.

Supplementary Methods

Patient populations

The UPMC Hillman Cancer Center registry was queried for unresectable stage III/IV melanoma from 2015-2020, who also received anti-PD1 (pembrolizumab or nivolumab; PD1 cohort), anti-PD1+anti-CTLA4 (nivolumab plus ipilimumab; I+N cohort), or BRAF±MEK inhibitors (vemurafenib, dabrafenib, encorafenib ± cobimetinib, trametinib, binimetinib; BRAF cohort) treatment (adjuvant therapy were excluded) (**Table S1**). Patients treated with ipilimumab before pembrolizumab or nivolumab were excluded given evidence that anti-CTLA4 therapy impacts molecular characteristics of anti-PD1 response¹. A detailed workflow of patient selection is illustrated in **Fig. S1**. The study protocol was approved by The University of Pittsburgh Institutional Review Board (IRB)-approved protocol (Protocol No. STUDY20020107). All samples have written informed patient consent.

Clinical data annotation and therapy response

Demographic, clinical, treatment, and radiologic variables measured using RECIST were collected for each patient from electronic medical record (EMR) databases as per our IRB-approved protocol (**Table S1**). For each cohort (PD1, I+N, or BRAF), we aggregated patients across lines of therapy based on the assumption that radiomic features of treatment response/resistance would be similar². For each patient, tumor sizes and best overall response were evaluated by a board-certified radiologist (S.Z., N.B.) by Response Evaluation Criteria in Solid Tumors (RECIST) v1.1³. Clinical outcomes were categorized into two groups: disease control (DC; including complete response [CR], partial response [PR], stable disease [SD]) and progressive disease (PD). Patients with high-quality CT and/or MRI imaging with measurable disease available were included in radiomics analysis (**Fig. S1**). A maximum of two lesions per organ were measured following RECIST v1.1 guidelines³. An illustration of the full analysis workflow is provided in **Fig. S2**.

Organ-level weighted RECIST scores

Given that radiomics data were generated from scans of target lesions, we computed a weighted RECIST score at the organ level to directly link radiomics with response derived from target lesions. The formula of $RECIST_{weighted}$ is as follows:

(1) if there is only one target lesion at an organ site, $RECIST_{weighted} =$

$$\frac{\textit{lesion size at best response} - \textit{lesion size at baseline}}{\textit{lesion size at baseline}},$$

(2) if there are two target lesions at an organ site, $RECIST_{weighted} =$

$$\frac{(\textit{lesion 01 size at best response} - \textit{lesion 01 size at baseline}) + (\textit{lesion 02 size at best response} - \textit{lesion 02 size at baseline})}{\textit{lesion 01 size at baseline} + \textit{lesion 02 size at baseline}}$$

Subsequently, organ response categories were defined as PD if $RECIST_{weighted} \geq 20\%$, and DC if $RECIST_{weighted} < 20\%$.

Inter-organ heterogeneity analysis

To assess inter-organ heterogeneity, we computed the standard deviation (SD) of $RECIST_{weighted}$ scores across all organ sites per patient, and compared between patients with uniform progression, mixed response, and uniform DC using two-sided Wilcoxon rank-sum tests. One metastasis refers to all target lesions at the same organ site in the same patient. Tumors at different organ sites are considered different metastases. Only patients with two or more metastases were included in the analysis.

Intra-organ heterogeneity analysis

Within each organ site, a maximum of two lesions were measured. For all organ sites, the largest lesion at baseline was defined as 01 and second as 02. To assess intra-organ heterogeneity, we computed the tumor size change for lesions 01 and 02 separately as follows:

$$\textit{lesion 01 tumor size change} = \frac{\textit{lesion 01 size at best response} - \textit{lesion 01 size at baseline}}{\textit{lesion 01 size at baseline}}$$

$$\textit{lesion 02 tumor size change} = \frac{\textit{lesion 02 size at best response} - \textit{lesion 02 size at baseline}}{\textit{lesion 02 size at baseline}}$$

Subsequently, we compared tumor size changes of lesion 01 *versus* 02 within organ sites using two-sided Wilcoxon signed-rank tests. In addition, we computed the absolute value of the difference between lesion 01 and 02 tumor size changes $\Delta = | \textit{lesion 01 tumor size change} - \textit{lesion 02 tumor size change} |$, and compared the organ-specific Δ between the overall response groups PD and DC using two-sided Wilcoxon rank-sum tests after log₁₀ transformation.

Radiomics image processing and feature extraction

Individual target lesions underwent semi-automatic segmentation with 3D Slicer⁴ (v4.10.2) to create a volume of interest (VOI) for radiomics analysis from pretreatment contrast-enhanced CT scans and T1 pre-contrast, T1 post-contrast, and T2/FLAIR MRI sequences. Images were pre-processed using Nyul Intensity Normalization⁵, a technique used to standardize the intensity levels of medical images, ensuring consistency and comparability across different scans and imaging modalities. It adjusts the pixel values within a scan, aligning them with a predefined intensity range acquired from a reference scan. Radiomic features were extracted from scans using Python (v3.8) packages including NumPy (v1.2) and SimpleITK (v2.2.0).

From each extracted VOI, a total of 400 features were derived, consisting of 10 histogram-based descriptors and 390 second-order texture features (**Table S2**). The description of the first-order and second-order features is provided in **Table S3**. The second-order features comprise 13 Haralick features⁶, each computed across five gray levels and three rotation-invariant measures. The naming convention for second-order features is "(Gray-level Quantization)_(Feature Name)_(Measure used for Rotation Invariance)", where (Gray-level Quantization) denoted as "8,16,32,64,256", (Feature Name) denoted as "Haralick feature name", and (Measure used for Rotation Invariance) denoted as average (avg), variance (var), and range of the Grey-Level Co-occurrence Matrix (GLCM) features computed across four different angles to ensure their rotational invariance. Further details are described as follows.

The first-order features were computed to describe the distribution of voxel intensities within each Region of Interest (ROI). These included statistical metrics such as Minimum, Maximum, Mean, Standard Deviation, Percentiles (1, 5, 95, 99), Skewness, and Kurtosis.

The second-order texture features are derived from the GLCM, which characterizes the spatial relationship of pixel intensities within an image. GLCM captures the frequency of occurrence of pairs of pixel intensity values at given spatial displacements and directions, providing valuable information about texture patterns. Each second-order feature was derived from GLCMs computed with respect to four angles (0°, 45°, 90°, and 135°), representing the angles between the pixel of interest and its neighboring pixels, allowing for a comprehensive capture of variations from various perspectives.

To ensure rotation invariance, statistical measures including average, mean, and range were computed across all angles for each feature resulting in 60 rotation-invariant features. These measures effectively encapsulate the feature's essence, regardless of its orientation. To enhance the signal-to-noise ratio, the original images underwent discretization into five grey levels (8, 16, 32, 64, and 256 grey levels), with 60 rotation-invariant features computed at each grey level. Furthermore, the set of 195 features is further refined by dividing each feature by the VOI, resulting in another set of 195 volume-independent features, leaving 400 radiomics features for subsequent analysis.

For MR imaging, preprocessing steps were undertaken per Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB) Brain Extraction Tool (BET)⁷ resulting in three distinct imaging phenotypes free of non-brain tissue: edema/invasion, active enhancing tumor and necrosis. Contralateral physiologic white matter was also segmented for within-sequence normalization⁸ and hemorrhage, and intracranial vasculature appearing as hyperintensity on pre-contrast T1W1 images were subtracted from FLAIR sequence images to prevent obscuration of tumor textural features.

Machine-learning model construction

Machine learning (ML) to predict overall and organ-specific responses using radiomic features was carried out using eXtreme Gradient Boosting⁹ (XGBoost). ML construction was performed using R package *caret*¹⁰ (v6.0.91), with steps, functions, and parameters described below.

To develop overall response DC/PD models: We constructed XGBoost models to predict overall response in I+N cohort ($n=82$) or PD1 cohort ($n=129$). Data were randomly split into training and test sets by an 80/20 ratio using function *createDataPartition*. In the training set, features were assessed to

exclude the ones with low variance, high collinearity, or strong correlation (Spearman's correlation $\rho > 0.8$). This step reduced 400 radiomics features to 17 for I+N models and 23 for PD1 models. The selected features are listed in **Fig. S7**. Subsequently, model optimization and hyperparameter tuning (`xgb_tune`) were performed with 10-fold cross-validation (CV) using function `train`, with parameters [`method = "xgbTree"`, `trControl = cv_opts`, `tuneGrid = xgb.grid`, `metric = "ROC"`]. `cv_opts` is the CV object generated by function `trainControl` with parameters [`method = "cv"`, `summaryFunction = twoClassSummary`, `number = 10`, `classProbs = TRUE`], and `xgb.grid` is the object generated by function `expand.grid` with parameters [`nrounds = 1000`, `eta = c(0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5)`, `max_depth = c(2, 3, 4, 5, 6, 7, 8, 9, 10)`, `gamma = c(1, 2, 3)`, `subsample = c(0.5, 0.75, 1)`, `min_child_weight = c(1, 2, 3)`, `colsample_bytree = c(1)`], representing various combinations of hyperparameters. The model of the highest AUC was selected as the final model. The model's performance on unseen data was evaluated in the test set using function `predict` with default parameters.

To develop organ-specific DC/PD models: We constructed XGBoost models to predict organ-specific response in I+N or PD1 cohort, including lung, lymph node [LN], liver, soft tissue (with CT features), and brain (with MRI features). The sample sizes varied across organs and cohorts: in I+N cohort, we had lung ($n=34$), LN ($n=37$), liver ($n=21$), soft tissue ($n=32$), and brain ($n=20$), while in PD1 cohort, we had lung ($n=54$), LN ($n=52$), liver ($n=22$), soft tissue ($n=50$). Due to the small sample size (<20) in the PD1 cohort for brain samples, we did not build models for brain responses in this cohort.

Recognizing that sample sizes for organ models are generally smaller than patient cohort sizes available for response prediction, and to ensure our methods were consistent for all organ models, we used a different approach from above. Data were not split into training and test sets; model's performance was evaluated through the Leave-One-Out CV (LOOCV) method, recognizing that the generalization of the models will require independent validation. For each organ site, a bootstrap strategy was implemented to reduce the initiated set of 400 radiomics features per organ to a more manageable number prior to model training. Taking lung from I+N cohort as an example, this process includes three steps:

(1) We initially selected features that were abundant in the majority of the samples. Specifically, we computed the arithmetic mean of each feature across 34 lung metastases, $feat_mean(i)$, where i represents feature 1,2,3...400. Features with values higher than $feat_mean(i)$ in at least 30% of the samples were retained, resulting a reduction from 400 features to 245;

(2) Subsequently, we performed bootstrapping using 80% of the samples. This was done by randomly sampling with replacement from the original dataset 100 times, creating 100 sets of samples. For each set of bootstrapped samples, we conducted two-sided Wilcoxon rank-sum tests to identify features differentially abundant between groups based on clinical outcomes (DC and PD). We kept features that passed $P < 0.20$, which is a lenient threshold commonly used in the literature¹¹. Adjustment for multiple comparisons was not applied in this scenario;

(3) Finally, we ranked the features based on their relevance to the outcome by summarizing the number of sample sets in which they were identified as differentially abundant, and selected top 10 features for model training. The selected features are listed in **Fig. S8**.

We repeated this procedure for all organs in I+N or PD1 cohort. The training process was performed using *train* function with [*method* = "L00CV"] as the *trainControl* parameter and other parameters as described above. The model of the highest AUC was selected as the final model.

To develop pan-organ DC/PD models: In addition to individual organ models, we constructed XGBoost models to predict pan-organ response across all metastases in I+N or PD1 cohort, with organ site included as a covariate. Data were split into training and test sets, with a procedure ensuring that metastases from the same patient fell into the same set (either training or test) to prevent data leakage. We treated metastases from the same patient as a single block, assigning each block a unique index. Patient blocks were then split by an 80/20 ratio using function *createDataPartition*, and the original organ-based training and test sets were reconstructed based on these indexes.

In the training set, features were assessed to exclude the ones with low variance, high collinearity, or strong correlation (Spearman's correlation $\rho > 0.8$). This step reduced 400 radiomics features to 15 for I+N models and 24 for PD1 models. The selected features are listed in **Fig. S9**. Model optimization and hyperparameter tuning were the same as that of overall response models described

above. The model of the highest AUC was selected as the final model. The model's performance on unseen data was evaluated in the test set using function *predict* with default parameters.

ROC generation and variable importance evaluation

For overall response or pan-organ response DC/PD models, we reported two types of Receiver Operating Characteristic (ROC) curves for the final model. This includes: (1) the training set 10-fold CV ROC curve, representing the CV prediction by the final model during the training process. These ROC curves were generated using function *evalm* with parameters [*list1 = xgb_tune, positive = "DC"]* from R package MLevel (v0.3). *xgb_tune* is the XGBoost training object generated by function *train* from R package caret (v6.0.91), with detailed parameters described in the model construction methods section above; and (2) the test set ROC curve, representing the performance of the final model on unseen data. These ROC curves were generated using function *roc* from R package pROC (v1.18.0). For organ-specific response DC/PD models, we reported the LOOCV ROC curve for the final model using function *evalm*.

For all models, sensitivity and specificity were also reported alongside AUC metrics. The p-values for AUC were computed using function *roc.area* from R package verification (v1.42), which implements a two-sided Wilcoxon rank-sum test to determine statistical significance. Variable importance (VarImp) of each model was computed using function *varImp* from R package caret (v6.0.91) with parameter [*scale = TRUE*].

Comparison of DC and PD groups for detection of differential radiomic features

To independently evaluate the association between radiomics features and clinical outcomes, we performed statistical comparisons of DC and PD groups to directly detect differentially abundant radiomics features associated with the clinical outcome. Groups of at least five samples were included in this analysis. All data were first shifted to non-negative values by adding a constant, followed by log10 transformation. Starting with 400 radiomics features, we removed features that did not exceed their mean value in at least 30% of the samples prior to statistical analysis. Differentially abundant features were

identified using two-sided Wilcoxon rank-sum tests. A FDR-adjusted p-value threshold of <0.10 was applied to comparisons between overall response categories (DC *versus* PD), and a nominal p-value threshold of <0.01 was applied to organ response comparisons.

Statistical analysis

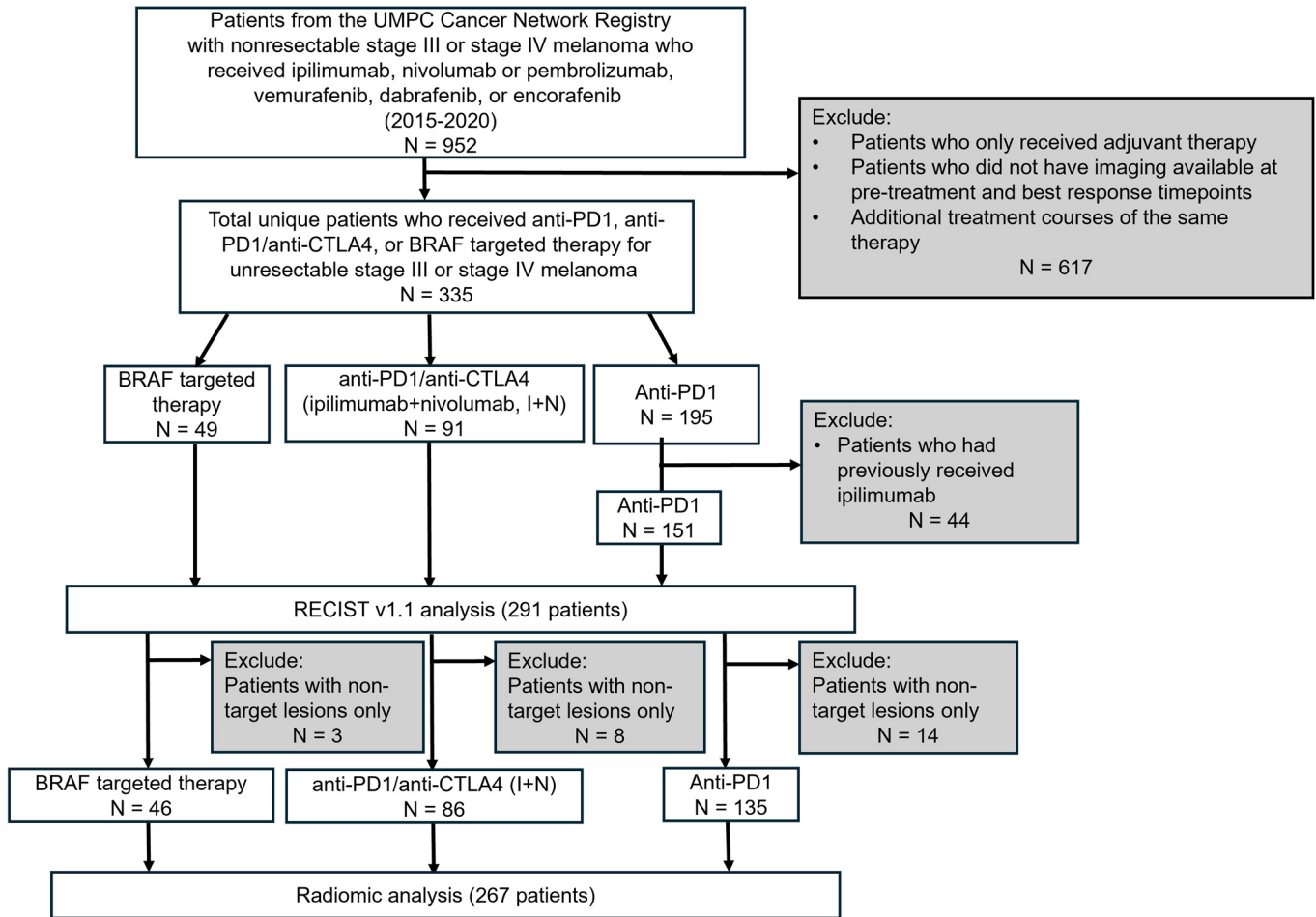
Among the clinical variables, patients with high/low neutrophil to lymphocyte ratio (NLR) were stratified by $NLR \geq$ or < 3.0 , and those with high/low lactate dehydrogenase (LDH) were stratified by $LDH \geq$ or < 170 , per literature review for high-risk status and institutional laboratory standards for patients with metastatic melanoma^{12,13}. The association between mixed organ responses in patients and PD as the overall response was assessed by Fisher's exact test. SD of metastasis site's $RECIST_{weighted}$ scores was compared between patients with uniform PD, mixed responses, and uniform DC by Wilcoxon rank-sum test. Metastasis site's $RECIST_{weighted}$ scores were compared between overall response groups DC/PD using Wilcoxon rank-sum test. The absolute differences in lesions' tumor size change% in lesion 01 *versus* 02 were log₁₀-transformed and then compared between overall response groups DC/PD using Wilcoxon rank-sum test. The statistical significance of lesion's tumor size change% in lesion 01 *versus* 02 was determined by Wilcoxon signed-rank test. Metastasis site's $RECIST_{weighted}$ scores were compared between patients with cutaneous *versus* non-cutaneous melanoma using Wilcoxon rank-sum test. Differentially abundant radiomics features between DC and PD were identified using Wilcoxon rank-sum test. Other statistical tests were described in the relevant method sections above. FDR was controlled at 0.10. Benjamini-Hochberg (BH)-FDR procedure¹⁴ was used for multiple comparison adjustments. All tests were two-sided.

References for Supplementary Methods

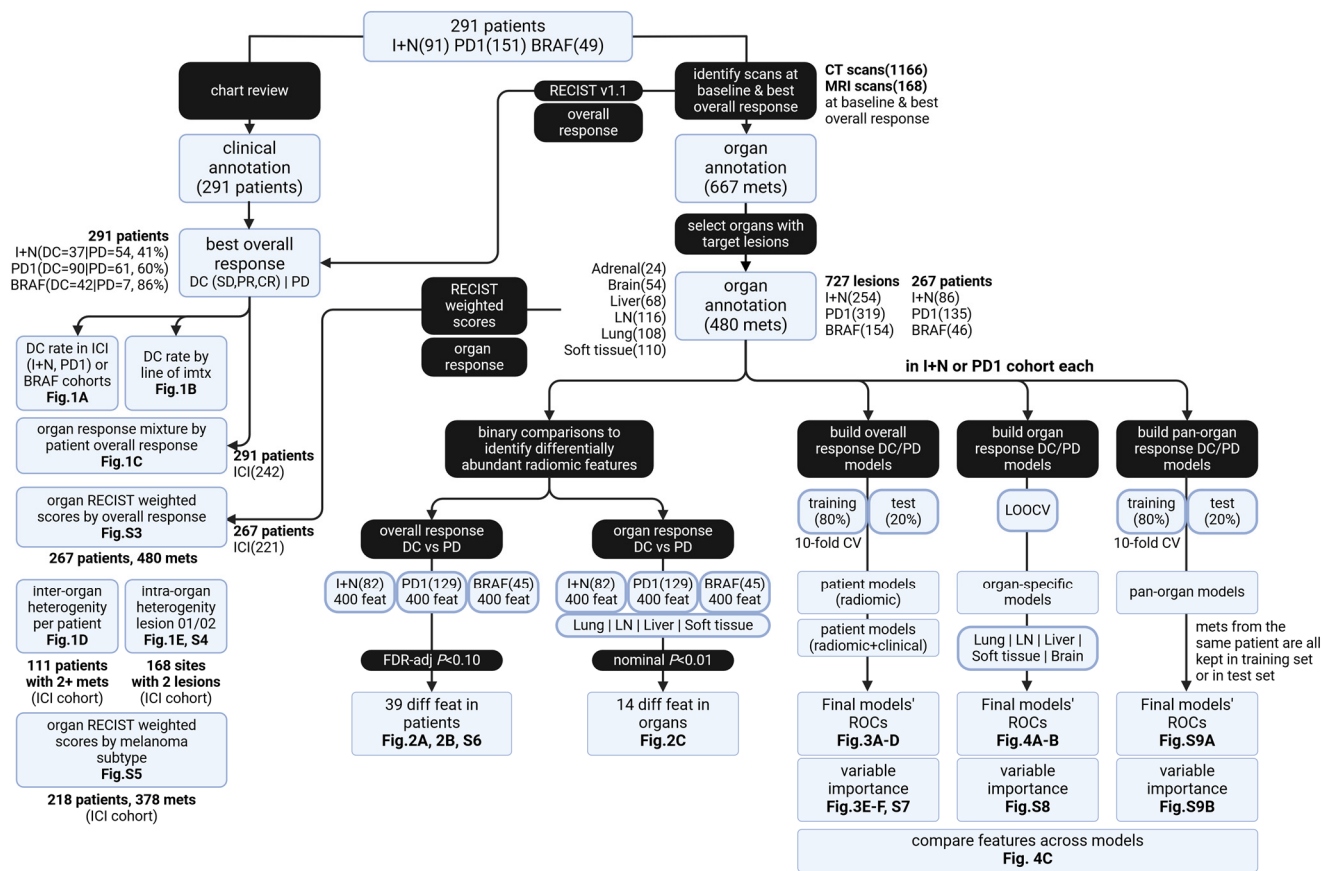
1. Campbell KM, Amouzgar M, Pfeiffer SM, et al. Prior anti-CTLA-4 therapy impacts molecular characteristics associated with anti-PD-1 response in advanced melanoma. *Cancer Cell* 2023;41(4):791-806 e4. doi: 10.1016/j.ccell.2023.03.010
2. Olson DJ, Eroglu Z, Brockstein B, et al. Pembrolizumab Plus Ipilimumab Following Anti-PD-1/L1 Failure in Melanoma. *J Clin Oncol* 2021;39(24):2647-55. doi: 10.1200/JCO.21.00079 [published Online First: 2021/05/05]
3. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009;45(2):228-47. doi: 10.1016/j.ejca.2008.10.026 [published Online First: 2008/12/23]
4. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging* 2012;30(9):1323-41. doi: 10.1016/j.mri.2012.05.001 [published Online First: 2012/07/06]
5. Nyul LG, Udupa JK, Xuan Z. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging* 2000;19(2):143-50. doi: 10.1109/42.836373
6. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics* 1973;SMC-3(6):610-21. doi: 10.1109/TSMC.1973.4309314
7. Zinn PO, Singh SK, Kotrotsou A, et al. A Coclinal Radiogenomic Validation Study: Conserved Magnetic Resonance Radiomic Appearance of Periostin-Expressing Glioblastoma in Patients and Xenograft Models. *Clin Cancer Res* 2018;24(24):6288-99. doi: 10.1158/1078-0432.CCR-17-3420 [published Online First: 2018/07/29]
8. Nyul LG, Udupa JK, Zhang X. New variants of a method of MRI scale standardization. *IEEE Trans Med Imaging* 2000;19(2):143-50. doi: 10.1109/42.836373
9. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery, 2016:785–94.

10. Kuhn M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 2008;28(5):1 - 26. doi: 10.18637/jss.v028.i05
11. Herridge MS, Cheung AM, Tansey CM, et al. One-year outcomes in survivors of the acute respiratory distress syndrome. *N Engl J Med* 2003;348(8):683-93. doi: 10.1056/NEJMoa022450
12. Cohen JT, Miner TJ, Vezeridis MP. Is the neutrophil-to-lymphocyte ratio a useful prognostic indicator in melanoma patients? *Melanoma Manag* 2020;7(3):MMT47. doi: 10.2217/mmt-2020-0006
[published Online First: 20200825]
13. Tumeh PC, Hellmann MD, Hamid O, et al. Liver Metastasis and Treatment Outcome with Anti-PD-1 Monoclonal Antibody in Patients with Melanoma and NSCLC. *Cancer Immunol Res* 2017;5(5):417-24. doi: 10.1158/2326-6066.CIR-16-0325 [published Online First: 20170414]
14. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* 1995;57(1):289-300.

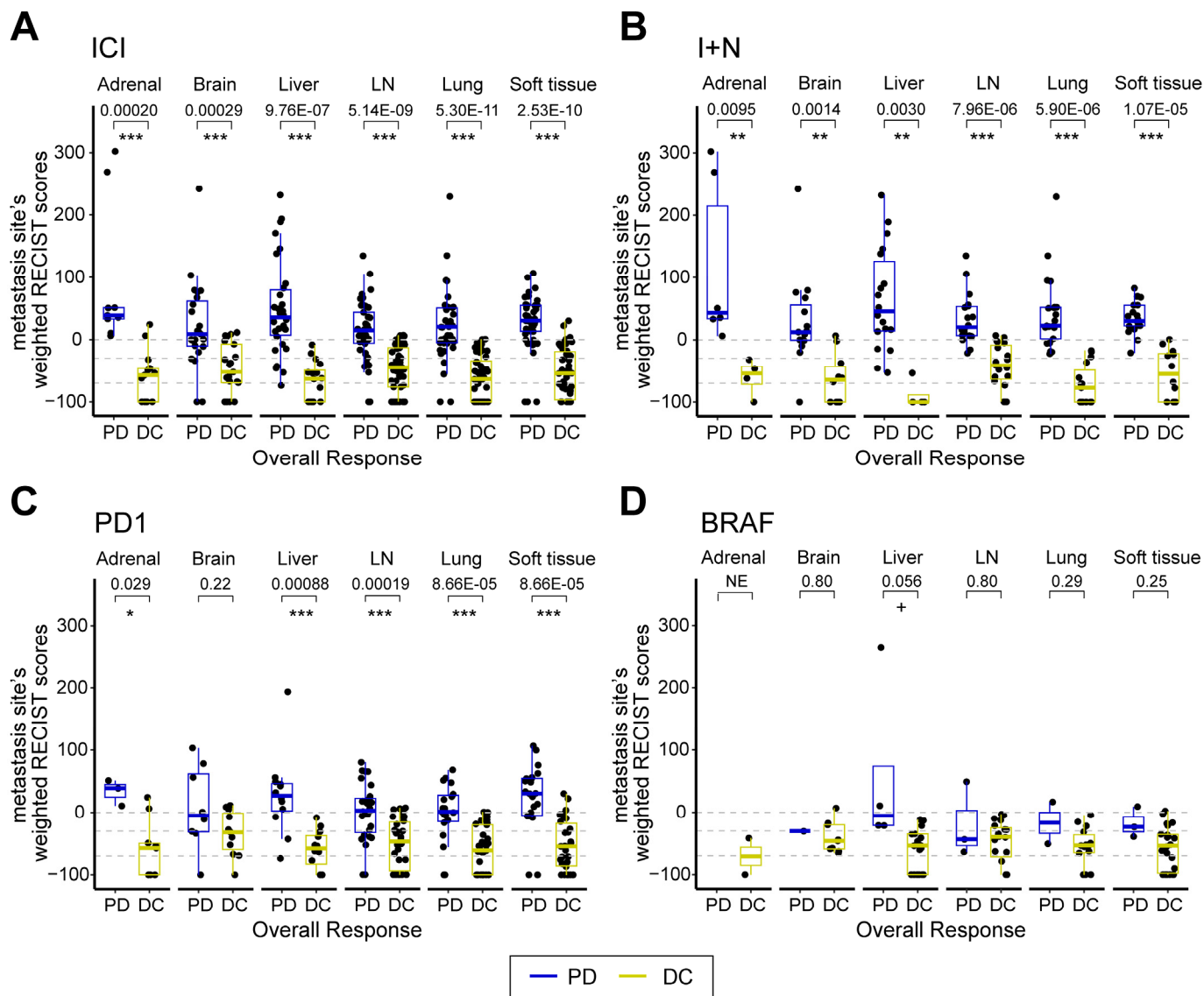
Supplementary Figures



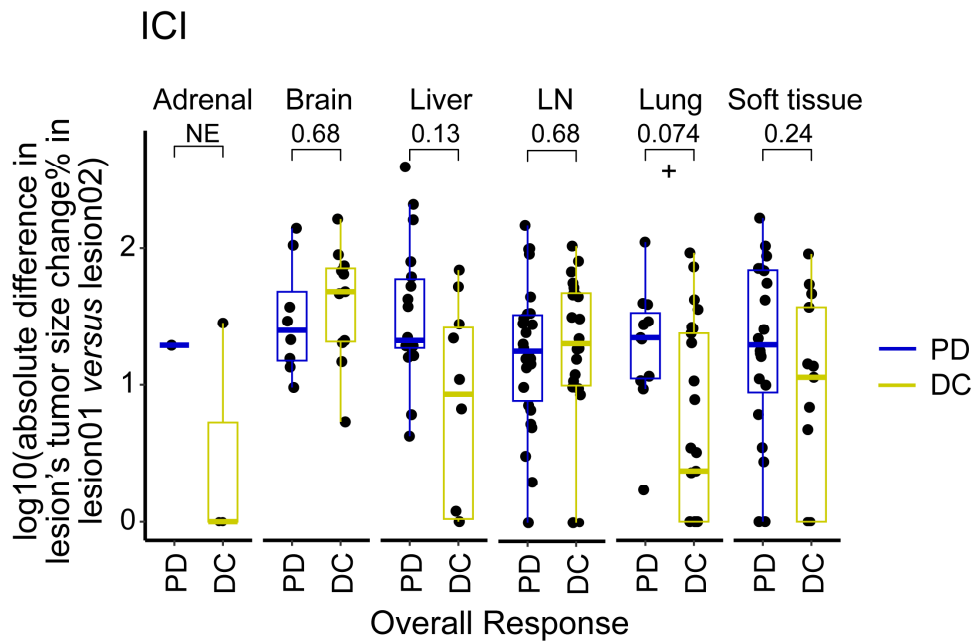
Supplemental Figure 1. Workflow for identification of analysis population. Patient selection process includes: (1) Identify patients who received ipilimumab/nivolumab (I+N), PD1 monotherapy, or BRAF targeted therapy and have appropriate imaging available; (2) Exclude patients who received therapy as an adjuvant treatment; (3) If a patient received the same therapy for multiple separate treatment courses only include in each cohort once, with a preference for earlier treatment course if appropriate imaging is available; and (4) In PD1 cohort exclude patients who had received ipilimumab prior to PD1 treatment.



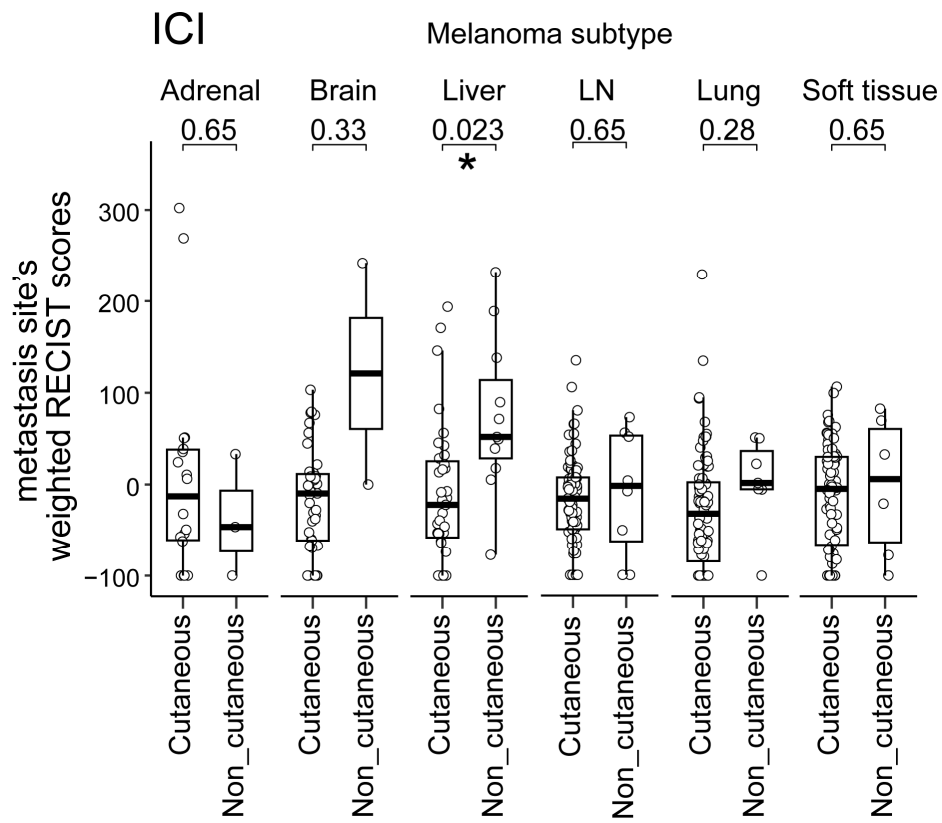
Supplemental Figure 2. Overall analysis workflow. Clinical data and medical images from 291 patients were used for the analysis of overall and organ-specific response, inter- and intra-organ heterogeneity, differentially abundant radiomics features in DC *versus* PD, and machine-learning models predicting overall response or organ-specific response or organ-specific response. Multiple lesions at the same organ site are considered one metastasis. Mets = metastases. Imtx = immunotherapy. LN = lymph node. CV = cross-validation. LOOCV = leave-one-out CV. ROC = Receiver Operating Characteristic.



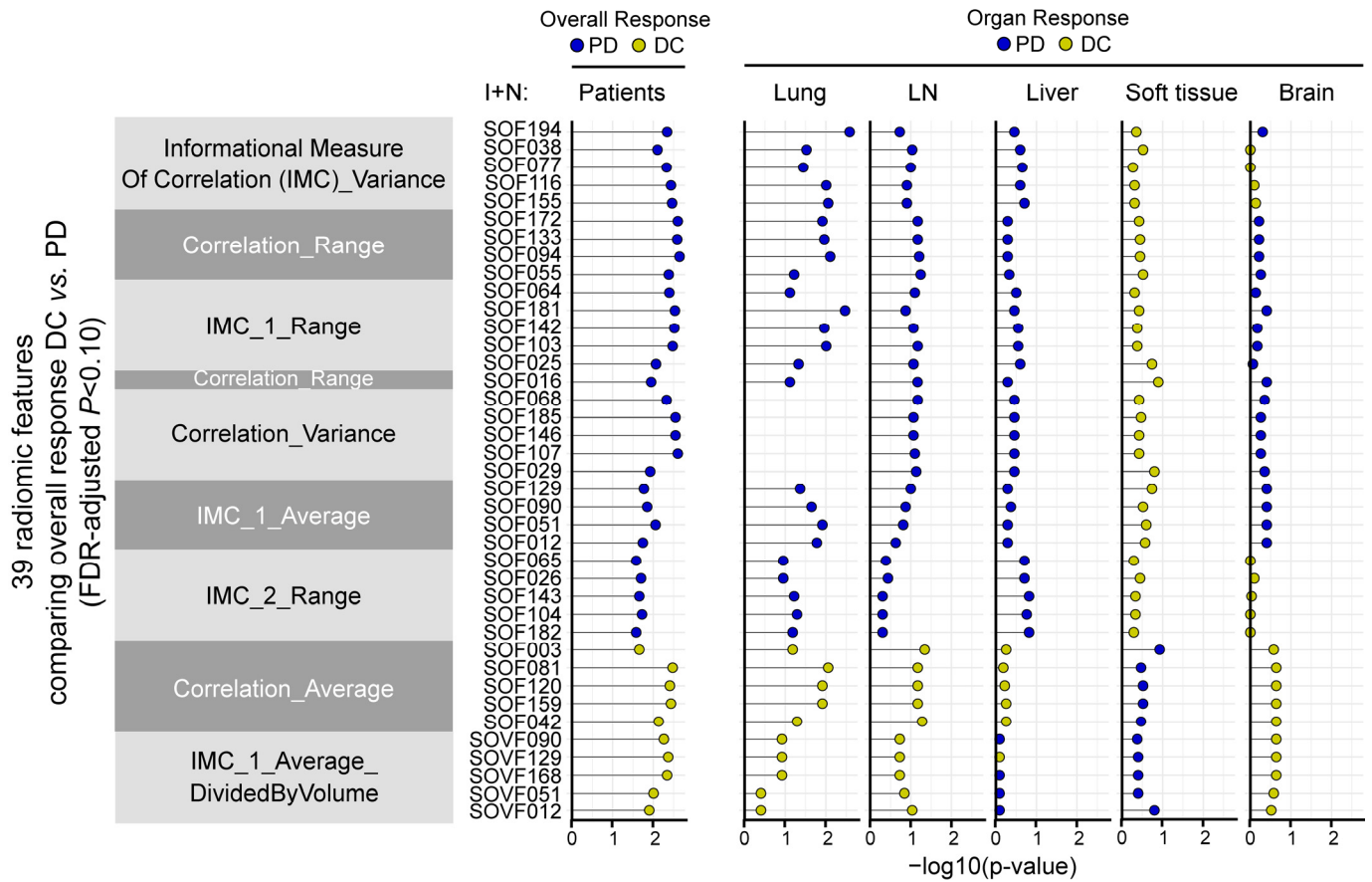
Supplemental Figure 3. Organ's weighted RECIST scores comparing overall response differ by organ site. (A) ICI cohort. (B) I+N cohort. (C) PD1 cohort. (D) BRAF cohort. Each data point represents one metastasis at each organ site. $n=480$ metastases from 267 patients are shown. LN = lymph node. Two-sided Wilcoxon rank-sum test was used in **A-D**. FDR-adjusted p-values are shown. FDR was controlled at 0.10. Denotations: *** $P<0.001$, ** $P<0.01$, * $P<0.05$, + $P<0.10$.



Supplemental Figure 4. Intra-organ heterogeneity by overall response in ICI cohort. The y-axis represents log₁₀-transformed absolute differences in (lesion01's tumor size change% minus lesion 02's tumor size change%). $n=168$ sites from ICI patients who had two lesions per organ site are shown. LN = lymph node. Two-sided Wilcoxon rank-sum test was used. FDR-adjusted p-values are shown. FDR was controlled at 0.10. Denotations: + $P < 0.10$.

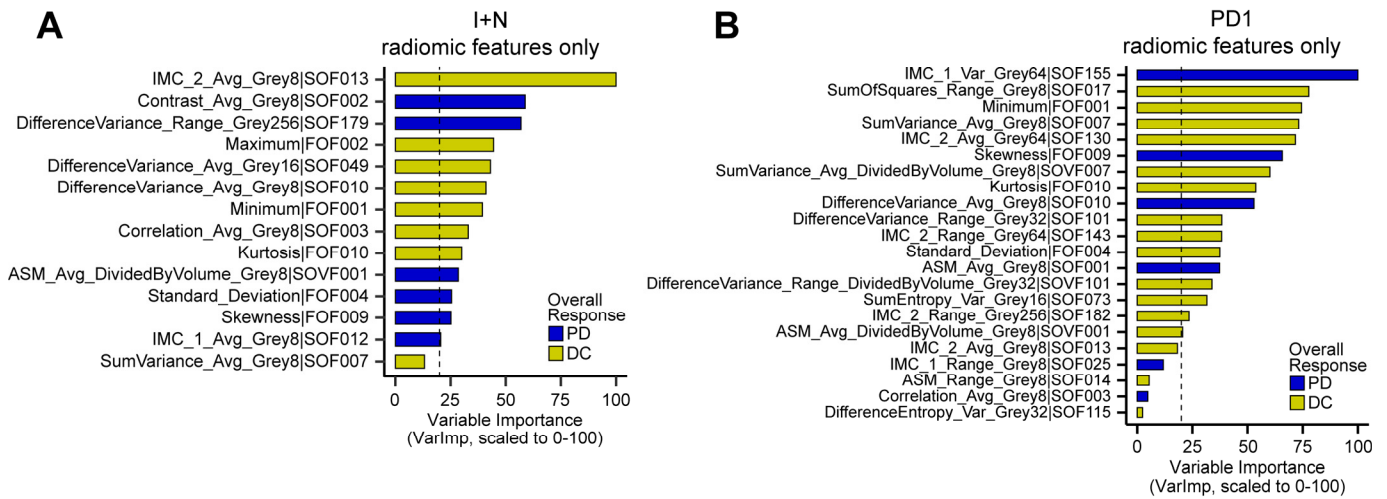


Supplemental Figure 6. Organ's weighted RECIST scores by melanoma subtype in ICI cohort. LN = lymph node. Two-sided Wilcoxon rank-sum test was used. FDR-adjusted p-values are shown. FDR was controlled at 0.10. Denotations: * $P < 0.05$.

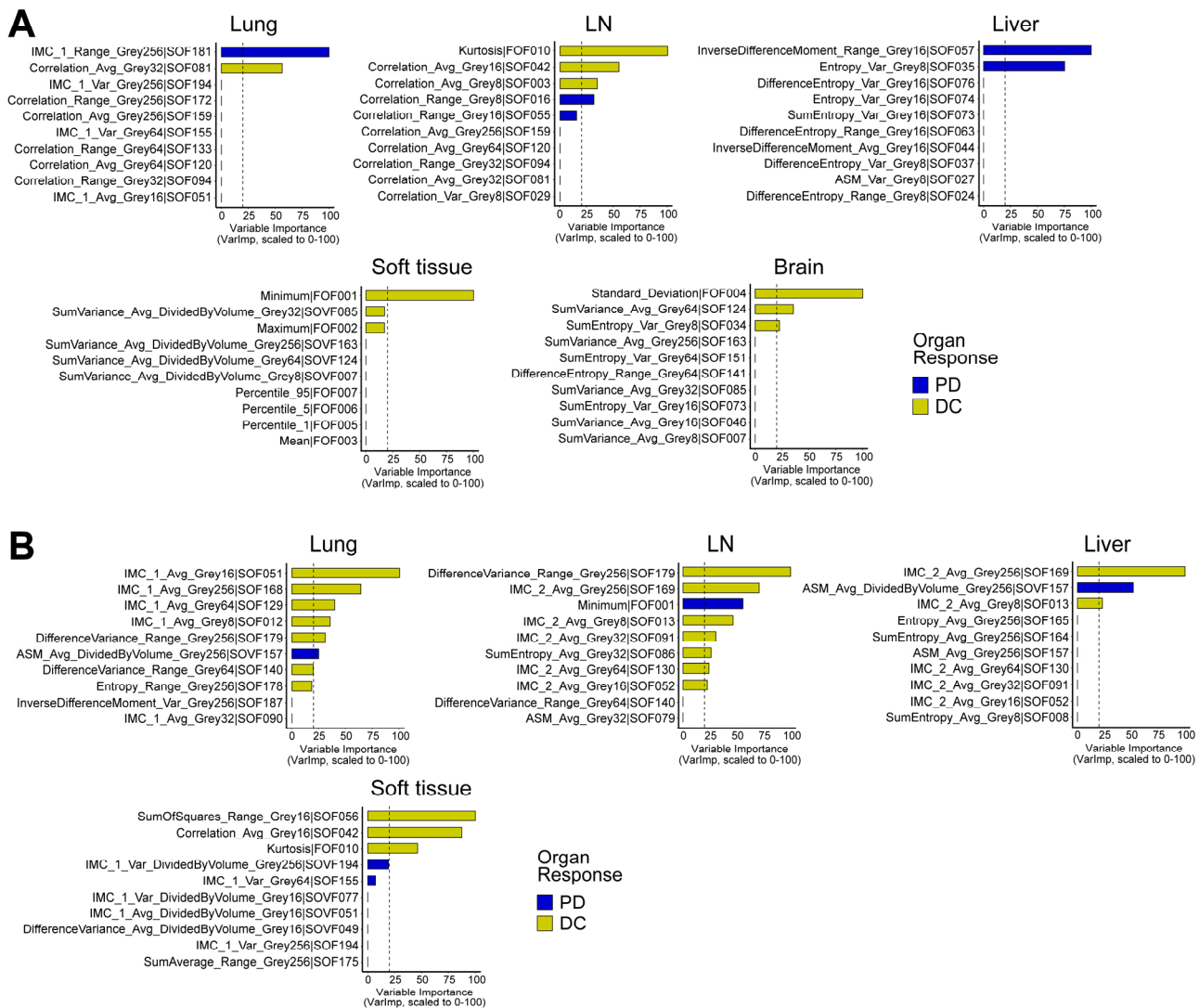


Supplemental Figure 6. The distribution of the 39 radiomics features that distinguish overall response DC/PD in I+N cohort by organ site and with organ response colored by DC or PD.

Features are shown in the same order as on the heatmap from **Fig. 2A** (FDR-adjusted $P < 0.10$). Color indicates whether a feature is greater in organ response PD (blue) or DC (gold). LN = lymph node. IMC = Informational Measure of Correlation. Full feature IDs and names are described in **Tables S2** and **S3**. Two-sided Wilcoxon rank-sum test was used.



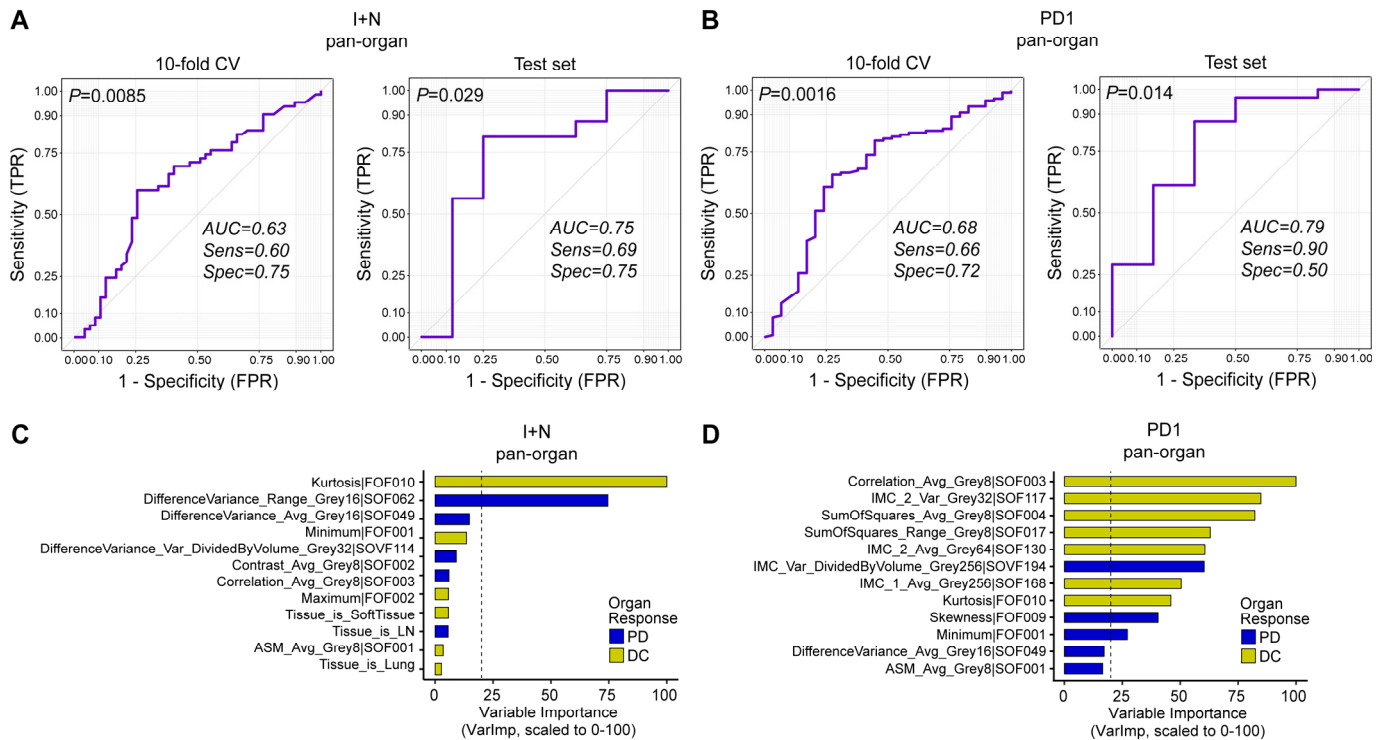
Supplemental Figure 7. Variable importance (VarImp) of the radiomics only models predicting overall response DC/PD in I+N or PD1 cohort. (A) I+N cohort, corresponding to the model in Fig. 3A. (B) PD1 cohort, corresponding to the model in Fig. 3B. Features with VarImp >1 are shown; red vertical dashed line indicates VarImp=20; features with VarImp ≥20 are generally considered important in predicting outcome. Color indicates whether a feature is greater in overall response PD (blue) or DC (gold). IMC = Informational Measure of Correlation. ASM = Angular Second Moment.



I+N

PD1

Supplemental Figure 8. Variable importance of the radiomics models predicting organ-specific response DC/PD at each organ site from I+N or PD1 cohort. (A) I+N cohort, corresponding to the models in Fig. 4A. (B) PD1 cohort, corresponding to the models in Fig. 4B. In each model, the 10 features used for model training are shown; red vertical dashed line indicates VarImp=20; features with VarImp \geq 20 are generally considered important in predicting outcome. Color indicates whether a feature is greater in organ response PD (blue) or DC (gold). LN = lymph node. IMC = Informational Measure of Correlation. ASM = Angular Second Moment.



Supplemental Figure 9. Pan-organ models predicting organ-specific response DC/PD in I+N or PD1 cohort. Pan-organ models were developed across all metastases, with organ-organ site as a covariate. We confirmed that individual patients' organ metastases were either all in training or test set to prevent data leaking. For each cohort, models were optimized in the training set with 10-fold CV, and the final performance was reported on unseen data in the test set. We show both the training set 10-fold CV ROC curve as well as the test set ROC curve. AUC, Sensitivity (Sens), and Specificity (Spec) were reported. **(A)** Model of radiomic features in I+N cohort. $n=109$ and 24 metastases in training/test set (80% / 20% split), respectively (total is 133). 400 radiomic features were reduced to 15 prior to model training. **(B)** Model of radiomic features in PD1 cohort. $n=152$ and 37 patients in training/test set (80% / 20% split), respectively (total is 189). 400 radiomic features were reduced to 24 prior to model training. **(C)** Variable importance (VarImp) of the features from I+N model in **A**. **(D)** Variable importance (VarImp) of the features from PD1 model in **B**. Features with $\text{VarImp} > 1$ are shown in **C** and **D**; red vertical dashed line indicates $\text{VarImp}=20$; features with $\text{VarImp} \geq 20$ are generally considered important in predicting outcome. Color indicates whether a feature is greater in organ response PD (blue) or DC (gold). ROC = Receiver Operating Characteristic. AUC = Area Under Curve. CV = cross-validation. FPR = false positive rate. TPR = true positive rate. IMC = Informational Measure of Correlation. ASM = Angular Second

Moment. The AUC p-value shown at the top left corner of each ROC panel in **A-B** was computed using function *roc.area* from R package *verification* (v1.42) which implements a two-sided Wilcoxon rank-sum test.

Supplementary Tables

Supplemental Table 1. Demographic and clinical characteristics of the patients.

Supplemental Table 2. List of 400 radiomics features extracted from scans.

Supplemental Table 3. Description and definition of the radiomics features.

Supplemental Table 4. Tumor size changes per organ per cohort as shown in Fig. S3.

Supplemental Table 5. Comparison of radiomics features across models as shown in Fig. 4C.